DOES LLM PRE-TRAINING TYPICALLY OCCUR AT THE EDGE OF STABILITY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Quadratic approximations provide a common lens for studying neural network optimization, but recent evidence challenges their predictive validity. In full-batch gradient descent with learning rate (LR) η, Cohen et al. (2021) observed the Edge of Stability (EoS), where the largest Hessian eigenvalue concentrates near $2/\eta$, seemingly at odds with classical stability conditions predicted by quadratic approximations. In this work, we revisit the fidelity of quadratic approximations as a model of neural network training dynamics, with a particular focus on their failure modes in LLM training. We first highlight a distinct failure mode of the quadratic approximation, where persistent negative curvature causes the quadratic dynamics to diverge regardless of the LR, even though the real dynamics remain stable. We term this phenomenon the *Edge of Convexity* (EoC). After removing the effect of this LR-independent instability, we then propose a generalized definition of *Edge of Stability* (EoS) for stochastic training with adaptive optimizers, based on whether increasing the LR induces significant divergence in the quadratic dynamics while leaving the real dynamics largely unchanged. Across different LLM pretraining runs with various model sizes up to 787M, we find that (1) EoC is always observed across LLM pretraining, and (2) EoS is also prevalent but not universal: it disappears when the LR becomes sufficiently small (e.g., after decay) or when the batch size falls below a critical threshold that has a positive correlation with the critical batch size. Together, these findings characterize when and how quadratic approximations fail and provide a foundation for future work on understanding training dynamics.

1 INTRODUCTION

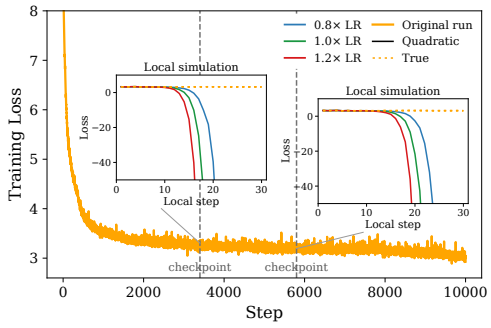
In optimization theory, the dynamics of gradient-based methods are commonly analyzed through a local *quadratic approximation* of the loss function, given by the second-order Taylor expansion around the current iterate:

$$\mathcal{L}(\boldsymbol{w}_{t+1}) \approx \mathcal{L}(\boldsymbol{w}_t) + (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \nabla \mathcal{L}(\boldsymbol{w}_t) + \frac{1}{2}(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)^\top \nabla^2 \mathcal{L}(\boldsymbol{w}_t)(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t). \quad (1)$$

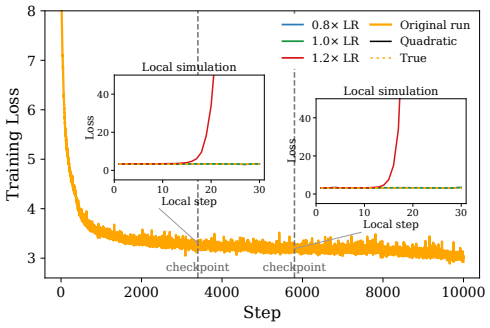
This approximation yields the well-known *descent lemma* for gradient descent, which guarantees a decrease in loss when the LR η satisfies $\eta < 2/\lambda_{\max}$. Here, $\lambda_{\max}$ is the maximum eigenvalue of the Hessian, which serves as a measure of the local curvature of the loss function. A standard analysis pipeline in optimization theory then proceeds by bounding quantities that appear in the lemma and deriving guarantees on the loss convergence.

However, in deep learning, the quadratic approximation often fails to accurately capture the true training dynamics. Notably, Cohen et al. (2021) show that full-batch gradient descent (GD) typically operates at the *Edge of Stability* (EoS) when training neural networks. For a fixed LR η, the maximum Hessian eigenvalue $\lambda_{\max}$ tends to increase during training and eventually oscillates around $2/\eta$, thus $\eta \approx 2/\lambda_{\max}$. Under the quadratic approximation, this places the dynamics exactly at the stability boundary of gradient descent, since the quadratic model predicts loss divergence once $\eta > 2/\lambda_{\max}$. Consequently, the descent lemma can no longer guarantee a persistent decrease of the loss. Yet in practice, the loss continues to decrease over long time scales. Similar EoS behavior has also been observed for full-batch adaptive optimizers such as Adam and RMSProp (Cohen et al., 2022), when curvature is measured using an appropriately preconditioned Hessian.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Illustration of Edge of Convexity (EoC).



(b) Illustration of Edge of Stability (EoS).

Figure 1: LLM pretraining on FineWeb-10B exhibits two distinct failure modes of the quadratic approximation. (a) **EoC phenomenon**: regardless of the LR choice, the training dynamics under quadratic approximation diverge due to persistent negative curvature, while the true dynamics remain stable. (b) **EoS phenomenon**: after compensating for negative curvature, the quadratic loss remains as stable as the true loss for $0.8\times$ and $1.0\times$ the training LR, but increasing the LR slightly to $1.2\times$ causes the quadratic loss to diverge. This shows that the current training LR is close to a critical stability threshold. See Sections 4.1 and 4.2 for details.

The EoS phenomenon therefore exposes a fundamental limitation of existing optimization analyses for deep learning, as many such analyses critically rely on the descent lemma and the underlying quadratic approximation. This discrepancy has motivated a line of work that seeks to develop tools to understand training dynamics without critically relying on the quadratic approximation (Arora et al., 2022; Damian et al., 2022; Cohen et al., 2024).

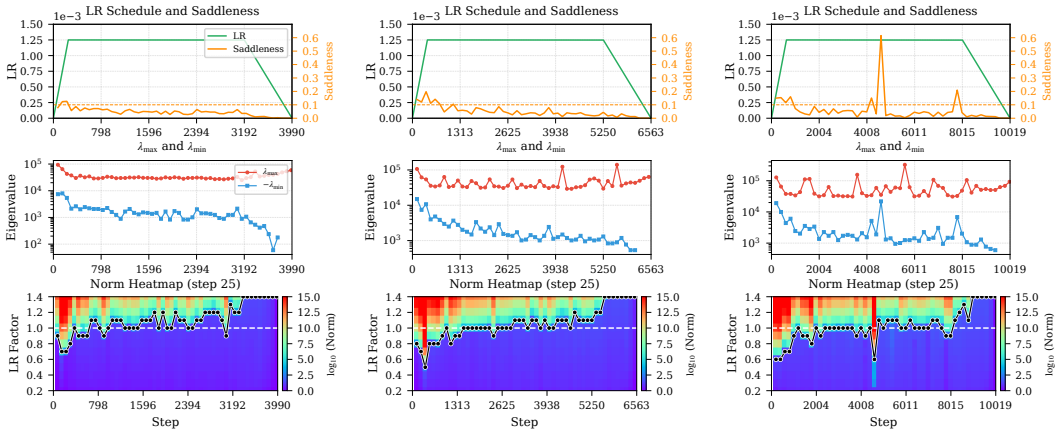
Despite these advances, existing empirical observations and theoretical analyses remain limited in several important aspects: (i) Most theoretical definitions of EoS focus on full-batch training, whereas stochastic training is more common in practice; (ii) Most studies focus on failures of the quadratic approximation caused by large positive Hessian eigenvalues, even though negative Hessian eigenvalues can also induce a similar gap between theory and practice, as we will elaborate in this work; (iii) Due to the prohibitive computational cost of computing and analyzing the Hessian, direct measurements of EoS are largely limited to relatively small-scale settings such as CIFAR-10 and ImageNet, leaving open whether the same failure mode persists in large-scale training.

Our Contributions. We present a systematic study of the failure modes of the quadratic approximation in stochastic, adaptive, and large-scale training, with a particular focus on LLM pretraining.

- We first highlight an LR-independent failure mode of the quadratic approximation, where persistent negative curvature during training causes the quadratic dynamics to diverge regardless of the LR, even though the real dynamics remain stable. We term this phenomenon the *Edge of Convexity* (EoC). See Section 4.1.
- We isolate and remove the effect of EoC, and then study the *Edge of Stability* (EoS) as an LR-dependent failure mode. We propose a generalized definition of EoS for stochastic training with adaptive optimizers, based on whether increasing the LR induces significant divergence in the quadratic dynamics while leaving the real dynamics largely unchanged. For full-batch GD, our definition recovers the original EoS definition. See Sections 4.2 and 4.3.
- We design efficient methods to detect both EoC and EoS at scale with theoretical guarantees (Appendices B.2 and B.3). Our results on LLM pretraining suggest that (1) EoC consistently arises; (2) EoS is prevalent but not universal, and disappears when the LR becomes sufficiently small (e.g., after decay) or when the batch size falls below a critical threshold that scales linearly with the critical batch size (Sections 5.2 to 5.4).

2 Preliminary

Notation. Throughout, $\|\cdot\|$ denotes the spectral norm for matrices and the ℓ_2 norm for vectors. We write $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ for the largest and smallest eigenvalues of a symmetric matrix $\boldsymbol{A}$, and $\rho(\boldsymbol{A})$ for the spectral radius of a square matrix $\boldsymbol{A}$. We write $\mathsf{vec}(\cdot)$ for vectorization and $\otimes$ for the Kronecker product. We use $\mathbb{S}_d$ for the set of $d \times d$ symmetric matrices and $\mathbb{S}_d^+$ for the subset of positive semidefinite matrices. For $\boldsymbol{S} \succeq 0$, we write $\|\boldsymbol{u}\|_{\boldsymbol{S}} = \sqrt{\boldsymbol{u}^\top \boldsymbol{S}\boldsymbol{u}}$. We use $[n]$ to denote $\{1, 2, \ldots, n\}$. We write $\nabla f(\boldsymbol{w})$ and $\nabla^2 f(\boldsymbol{w})$ for the gradient and Hessian of f at $\boldsymbol{w}$.



(a) GPT-small (104M). (b) GPT-medium (344M). (c) GPT-large (787M).

Figure 2: **EoS and EoC appear across model sizes.** At most checkpoints, the bottom heatmaps show a sharp transition from bounded displacement to explosion near LR factor 1.0, indicating that the training LR is at the critical threshold of EoS. In the middle row, $-\lambda_{\min}$ has a smaller magnitude than $\lambda_{\max}$, corresponding to saddleness below the 0.1 reference line in the top plots and consistent with EoC. Each heatmap column is a checkpoint, and the vertical axis gives the simulated LR divided by the training LR. See Section 5.2 for the results and Appendix C.1 for training and detection settings.

Problem Setup. We study a neural network $f(\boldsymbol{x};\boldsymbol{w}) : \mathbb{R}^{d+s} \to \mathbb{R}^q$, where $\boldsymbol{x} \in \mathbb{R}^d$ is the input and $\boldsymbol{w} \in \mathbb{R}^s$ is the trainable parameter. Given a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a per sample loss function ℓ, define the population loss and batch loss as $\mathcal{L}(\boldsymbol{w}) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell\big(f(\boldsymbol{x}_i;\boldsymbol{w}), y_i\big)$ and $\mathcal{L}_{\mathcal{B}}(\boldsymbol{w}) \coloneqq \frac{1}{B}\sum_{i\in\mathcal{B}} \ell\big(f(\boldsymbol{x}_i;\boldsymbol{w}), y_i\big)$, where $\mathcal{B} \subseteq [n]$ is a batch of size B.

Full-Batch Edge of Stability. We recall the full-batch Edge of Stability (EoS) introduced in Cohen et al. (2021). Consider full-batch gradient descent (GD) with LR η as $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta\nabla\mathcal{L}(\boldsymbol{w}_t)$. Let $\boldsymbol{H}_t \coloneqq \nabla^2\mathcal{L}(\boldsymbol{w}_t)$, and write $\lambda_{\max}(\boldsymbol{H}_t)$ for the maximal eigenvalue of $\boldsymbol{H}_t$. Classical local analysis says that, if the loss around $\boldsymbol{w}_t$ is well described by a positive semidefinite quadratic function, then GD is stable only when $\eta\lambda_{\max}(\boldsymbol{H}_t) \le 2$. Equivalently, the threshold $2/\eta$ is the largest curvature without causing instability.

Cohen et al. (2021) found that neural network training does not usually stay well below this threshold. During full batch training, $\lambda_{\max}(\boldsymbol{H}_t)$ often increases until it reaches a value near $2/\eta$. It then remains near this value for a long part of training. The loss is not monotone at every step, but it continues to decrease over longer time scales. They call this regime the *Edge of Stability* (EoS).

To relate this observation to our discussion on the fidelity of quadratic approximation, we fix a checkpoint $\boldsymbol{w}_{t_0}$ and consider

$$Q_{t_0}(\boldsymbol{w}) = \mathcal{L}(\boldsymbol{w}_{t_0}) + (\boldsymbol{w}-\boldsymbol{w}_{t_0})^\top\nabla\mathcal{L}(\boldsymbol{w}_{t_0}) + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_{t_0})^\top\boldsymbol{H}_{t_0}(\boldsymbol{w}-\boldsymbol{w}_{t_0}). \tag{2}$$

Suppose that this quadratic has a stationary point $\boldsymbol{w}_\star$ such that $\nabla\mathcal{L}(\boldsymbol{w}_{t_0}) + \boldsymbol{H}_{t_0}(\boldsymbol{w}_\star - \boldsymbol{w}_{t_0}) = 0$. Gradient descent on Q_{t_0} then satisfies

$$\boldsymbol{w}_{t+1} - \boldsymbol{w}_\star = (\boldsymbol{I} - \eta\boldsymbol{H}_{t_0})(\boldsymbol{w}_t - \boldsymbol{w}_\star).$$

If $\boldsymbol{H}_{t_0} \succeq 0$, this system is stable when $\rho(\boldsymbol{I} - \eta\boldsymbol{H}_{t_0}) \le 1$, or equivalently, $\eta\lambda_{\max}(\boldsymbol{H}_{t_0}) \le 2$. Once $\eta\lambda_{\max}(\boldsymbol{H}_{t_0}) > 2$, $\boldsymbol{w}_t$ oscillates with growing magnitude along a leading eigenvector of $\boldsymbol{H}_{t_0}$.

Based on this perspective, the observation that $\lambda_{\max}$ concentrates near $2/\eta$ throughout training is striking. It places training dynamics precisely at the boundary of stability predicted by the quadratic model, rather than safely within it. The EoS phenomenon therefore exposes a fundamental mismatch between the predictions of quadratic approximation and the actual behavior of neural network training, which is the starting point for our study of EoS in stochastic training.

3 Stability of Stochastic Gradient Descent on Quadratics

The full-batch EoS reviewed above is insightful but incomplete for modern neural networks since it relies on two simplifying conditions. First, the Hessian is assumed to be positive semidefinite. If

it has a negative eigenvalue, then the local quadratic dynamics contain an exploding direction for every positive LR, so the usual positive curvature stability threshold is no longer the only source of instability. Second, the analysis is deterministic and applies to full-batch gradient descent. It does not capture the additional randomness introduced by mini-batch sampling.

We then study these two effects on a stochastic quadratic objective. We consider a stochastic variant of Equation (2) to approximate the training loss. Specifically, at step t, let $f_t(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}_t \boldsymbol{x} + \boldsymbol{b}_t^\top \boldsymbol{x} + c_t$, where $\boldsymbol{A}_t$ is symmetric, $\mathbb{E}[\boldsymbol{A}_t] = \boldsymbol{A}$, and $\mathbb{E}[\boldsymbol{b}_t] = \boldsymbol{b}$. We assume that $\{(\boldsymbol{A}_t, \boldsymbol{b}_t)\}_{t\geq 0}$ are sampled independently from a fixed distribution, have finite second moments, and are independent of the initialization. SGD with LR η gives

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f_t(\boldsymbol{x}_t) = (\boldsymbol{I} - \eta \boldsymbol{A}_t)\boldsymbol{x}_t - \eta \boldsymbol{b}_t. \tag{3}$$

Let $\boldsymbol{M}_t \coloneqq \boldsymbol{I} - \eta \boldsymbol{A}_t$. Suppose that the population quadratic has a stationary point $\boldsymbol{x}_\star$ satisfying $\boldsymbol{A}\boldsymbol{x}_\star + \boldsymbol{b} = 0$. And we define $\boldsymbol{\epsilon}_t \coloneqq \boldsymbol{x}_t - \boldsymbol{x}_\star$, $\boldsymbol{r}_t \coloneqq \boldsymbol{A}_t \boldsymbol{x}_\star + \boldsymbol{b}_t$. Then $\mathbb{E}[\boldsymbol{r}_t] = 0$, the error $\boldsymbol{\epsilon}_t$ follows

$$\boldsymbol{\epsilon}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{A}_t)(\boldsymbol{x}_t - \boldsymbol{x}_\star) - \eta \boldsymbol{A}_t \boldsymbol{x}_\star - \eta \boldsymbol{b}_t = \boldsymbol{M}_t \boldsymbol{\epsilon}_t - \eta \boldsymbol{r}_t. \tag{4}$$

We study the mean square stability of the above stochastic quadratic dynamics.

Definition 3.1. For the stochastic dynamics in Equation (4), we say that the system is mean square stable if, for every initialization $\boldsymbol{\epsilon}_0 \in \mathbb{R}^d$ the resulting trajectory satisfies $\sup_{t\geq 0} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^2\right] < \infty$.

The next proposition gives the moment dynamics and a sufficient condition for mean square stability.

Proposition 3.2. *For each fixed η, there exist a matrix $\boldsymbol{S}$ and a vector $\boldsymbol{b}_{\text{quad}}$ that do not depend on t such that*

$$\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = (\boldsymbol{I} - \eta \boldsymbol{A})\mathbb{E}[\boldsymbol{\epsilon}_t], \tag{5}$$

$$\mathsf{vec}\big(\mathbb{E}[\boldsymbol{\epsilon}_{t+1}\boldsymbol{\epsilon}_{t+1}^\top]\big) = \mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]\mathsf{vec}\big(\mathbb{E}[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^\top]\big) + \boldsymbol{S}\mathbb{E}[\boldsymbol{\epsilon}_t] + \boldsymbol{b}_{\text{quad}}. \tag{6}$$

For every deterministic initialization, the affine system (4) has bounded second moments if $\rho(\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]) < 1$.

The proof is given in Appendix E.3. The above proposition implies that stochastic quadratics have two different instability modes. The first is caused by the negative curvature of the population quadratic and is independent of the LR threshold. The second is the usual positive-curvature instability, but in the stochastic setting its threshold is controlled by a second-moment operator rather than by $\lambda_{\max}(\boldsymbol{A})$ alone. The following arguments make the above ideas clear.

3.1 Negative Curvature Divergence

Suppose that $\lambda_{\min}(\boldsymbol{A}) < 0$, and let $\boldsymbol{u}_{\min}$ be a corresponding unit eigenvector. Then the negative curvature makes the stochastic quadratic dynamics unstable for every positive LR. Define the mean error in this direction by $a_t \coloneqq \boldsymbol{u}_{\min}^\top \mathbb{E}[\boldsymbol{\epsilon}_t]$. Using the mean dynamics $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}] = (\boldsymbol{I} - \eta \boldsymbol{A})\mathbb{E}[\boldsymbol{\epsilon}_t]$, we have

$$a_{t+1} = \boldsymbol{u}_{\min}^\top (\boldsymbol{I} - \eta \boldsymbol{A})\mathbb{E}[\boldsymbol{\epsilon}_t] = (1 - \eta \lambda_{\min}(\boldsymbol{A}))\, a_t = (1 + \eta |\lambda_{\min}(\boldsymbol{A})|)\, a_t.$$

Therefore, we have that $a_t = (1 + \eta|\lambda_{\min}(\boldsymbol{A})|)^t\, a_0$. For any initialization with $a_0 \neq 0$, this component grows without bound for every $\eta > 0$. Moreover,

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_t\|^2\right] \geq \|\mathbb{E}[\boldsymbol{\epsilon}_t]\|^2 \geq \left|\boldsymbol{u}_{\min}^\top \mathbb{E}[\boldsymbol{\epsilon}_t]\right|^2 = |a_t|^2.$$

Hence the second moment is also unbounded. This instability does not disappear when the LR is reduced. A smaller LR only slows down the escape along the negative curvature direction, while the quadratic model remains unstable for every positive LR.

3.2 Positive Curvature Divergence and Critical LR

In the remainder, we further see that even if we factor out those directions causing negative divergence and focus on the locally convex subspace, the instability can also arise from mini-batch noise and admits a critical LR threshold. To study the LR-dependent instability, we first need a convex approximation of the population quadratic that removes the LR-independent escape caused by negative curvature. The most direct construction is to discard the directions associated with negative

eigenvalues of $\boldsymbol{A}$ while leaving all nonnegative curvature directions unchanged. This gives the largest subspace on which the population quadratic is convex.

Let Π^+ be the orthogonal projection onto the eigenspace of $\boldsymbol{A}$ with positive eigenvalues. We apply this fixed projection at every step and define $\boldsymbol{M}_t^+ \coloneqq \Pi^+ \boldsymbol{M}_t \Pi^+$. The projected error dynamics are $\boldsymbol{\epsilon}_{t+1}^+ = \boldsymbol{M}_t^+ \boldsymbol{\epsilon}_t^+ - \eta \Pi^+ \boldsymbol{r}_t$. These dynamics keep the positive curvature part of the original stochastic quadratic while removing the negative curvature mode. We then define the critical LR for SGD using the mean square stability of this projected system.

Definition 3.3. For the stochastic quadratic dynamics in Equation (3), for every initialization $\boldsymbol{\epsilon}_0 \in \mathbb{R}^d$, define the critical LR on the nonnegative curvature subspace by

$$\eta_\star \coloneqq \sup \left\{ \eta > 0 : \sup_{t \geq 0} \mathbb{E}\left[\|\boldsymbol{\epsilon}_t^+\|^2 \right] < \infty \right\}. \tag{7}$$

For full batch GD, $\boldsymbol{A}_t \equiv \boldsymbol{A}$ and $\boldsymbol{b}_t \equiv \boldsymbol{b}$, so $\eta_\star = 2/\lambda_{\max}(\boldsymbol{A})$ when $\lambda_{\max}(\boldsymbol{A}) > 0$. The definition therefore recovers the full batch threshold in Section 2. For stochastic gradients, variation in $\boldsymbol{A}_t$ affects the second moment operator $\mathbb{E}[\boldsymbol{M}_t^+ \otimes \boldsymbol{M}_t^+]$, so the critical LR is not determined by $\lambda_{\max}(\boldsymbol{A})$ alone. Further stability analysis is given in Appendix D.2.

4 Stability of Stochastic Gradient Descent on Neural Networks

We now apply the stochastic quadratic analysis in Section 3 to neural network training with SGD. Fix a checkpoint $\boldsymbol{w}_{t_0}$. For a batch $\mathcal{B}_t$, we consider the quadratic surrogate

$$Q_{\mathcal{B}_t}(\boldsymbol{w}) = \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) + (\boldsymbol{w} - \boldsymbol{w}_{t_0})^\top g_{\mathcal{B}_t} + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{t_0})^\top \boldsymbol{H}_{\mathcal{B}_t}(\boldsymbol{w} - \boldsymbol{w}_{t_0}), \tag{8}$$

where $g_{\mathcal{B}_t} \coloneqq \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0})$ and $\boldsymbol{H}_{\mathcal{B}_t} \coloneqq \nabla^2 \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0})$. SGD with LR η on this quadratic is given by

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla Q_{\mathcal{B}_t}(\boldsymbol{w}_t). \tag{9}$$

As shown in Section 3, applying SGD to the local quadratic surrogate gives a stochastic linear system. Therefore, the two instability mechanisms identified for stochastic quadratics also arise in the local quadratic surrogate of neural network (NN) training in Equation (9).

First, we recall that $\boldsymbol{H}_{t_0} \coloneqq \nabla^2 \mathcal{L}(\boldsymbol{w}_{t_0}) = \mathbb{E}_{\mathcal{B}_t}[\boldsymbol{H}_{\mathcal{B}_t}]$ is the population Hessian at step t_0. If $\lambda_{\min}(\boldsymbol{H}_{t_0}) < 0$, the population quadratic $Q_t(\boldsymbol{w}) = \mathbb{E}[Q_{\mathcal{B}_t}(\boldsymbol{w})]$ is unbounded below as $t \to \infty$, and its mean dynamics have an expanding direction for every positive LR. This is the negative curvature failure studied in Section 3.1.

Second, we remove the negative curvature directions by projecting every local update. Let $\Pi_{t_0}^+$ project onto the positive eigenspace of $\boldsymbol{H}_{t_0}$, and let $\boldsymbol{\epsilon}_t^+$ denote the displacement of this projected run from the checkpoint. The update is

$$\boldsymbol{\epsilon}_{t+1}^+ = \Pi_{t_0}^+ \left[(\boldsymbol{I} - \eta \boldsymbol{H}_{\mathcal{B}_t}) \boldsymbol{\epsilon}_t^+ - \eta \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) \right], \qquad \boldsymbol{\epsilon}_0^+ \in \mathrm{range}(\Pi_{t_0}^+).$$

As in Definition 3.3, $\eta_\star(\boldsymbol{w}_{t_0})$ is the supremum of positive LRs for which $\sup_t \mathbb{E}\|\boldsymbol{\epsilon}_t^+\|^2 < \infty$ for every deterministic initial displacement in this subspace.

For NN training, once we replace the actual loss with $Q_{\mathcal{B}_t}(\boldsymbol{w})$, it is typically observed that the linear system formed by quadratic approximation diverges quickly, while the true dynamics of the NN remain stable (Figure 1a, Figure 1b). So a natural question is: which regime leads to the failure of quadratic approximation in NN training, negative or positive curvature divergence? We identify: NN training sits on an *edge* where either divergence mechanism is *just at the threshold* of breaking the local quadratic approximation.

4.1 Edge of Convexity

Negative curvature makes the quadratic model unstable. Negative Hessian eigenvalues in NNs have been widely reported in prior works (Sagun et al., 2016; 2017; Alain et al., 2019). In this paper, we show that they can make the quadratic model diverge while NN training remains stable (Figure 1a). If $\lambda_{\min}(\boldsymbol{H}_{t_0}) < 0$, SGD on the quadratic model in Equation (8) has an exploding direction for every $\eta > 0$ (Section 3.1). A smaller LR slows this explosion but cannot prevent it. We first measure the size of the negative curvature relative to the positive curvature at the same checkpoint.

Definition 4.1 (Saddleness). Suppose that $\lambda_{\max}(\boldsymbol{H}_{t_0}) > 0$. We define the saddleness at $\boldsymbol{w}_{t_0}$ by

$$\nu(\boldsymbol{w}_{t_0}) := \frac{\max\{0, -\lambda_{\min}(\boldsymbol{H}_{t_0})\}}{\lambda_{\max}(\boldsymbol{H}_{t_0})}. \tag{10}$$

This ratio measures the negative curvature relative to the largest positive curvature. For any symmetric matrix $\boldsymbol{H}$, $\inf_{\boldsymbol{H}_+ \succeq 0} \|\boldsymbol{H} - \boldsymbol{H}_+\| = \max\{0, -\lambda_{\min}(\boldsymbol{H})\}$. Thus the numerator in Equation (10) is the spectral norm distance from the Hessian to the positive semidefinite cone. Dividing by $\lambda_{\max}$ removes the overall scale of the Hessian. When ν is small, the Hessian is close to positive semidefinite relative to its largest positive curvature. It is still nonconvex whenever $\lambda_{\min} < 0$.

For the saddleness defined above, one observation is that for LLM pre-training, after the warm-up phase, $\nu \leq 0.1$, which leads to the first edge area of LLM pre-training.

Edge of Convexity (EoC) at $\boldsymbol{w}_{t_0}$. At a checkpoint $\boldsymbol{w}_{t_0}$, training is at EoC when $\lambda_{\min}(\boldsymbol{H}_{t_0}) < 0$, but saddleness $\nu(\boldsymbol{w}_{t_0})$ itself is small.

EoC therefore describes a particular failure of the quadratic approximation. The Hessian is close to positive semidefinite in relative scale, but its negative curvature is enough to make the quadratic model unstable for every positive LR.

Measuring population negative curvature during LLM pretraining. Early studies on small image and synthetic datasets reported negative Hessian eigenvalues much smaller in magnitude than the largest positive eigenvalue (Sagun et al., 2016; 2017; Alain et al., 2019). In this paper, we extend these measurements to the population Hessian throughout LLM pretraining (Figure 2). Doing so requires $\lambda_{\min}(\boldsymbol{H}_{t_0})$. The usual approach at scale is to run Lanczos on one mini-batch Hessian (Ghorbani et al., 2019; Yao et al., 2020; Zhang et al., 2024). This gives $\lambda_{\min}(\boldsymbol{H}_{\mathcal{B}})$. And since $\lambda_{\min}(\cdot)$ is concave, Jensen's inequality gives $\mathbb{E}[\lambda_{\min}(\boldsymbol{H}_{\mathcal{B}})] \leq \lambda_{\min}(\boldsymbol{H}_{t_0})$, so the mini-batch value is biased below the population value. This downward bias can be substantial in practice. To this end, we propose a stochastic power iteration for the estimate of $\lambda_{\min}(\boldsymbol{H}_{t_0})$, which further gives the systematic detection of EoC. The complete EoC detection procedure is given in Appendix B.2, along with the theoretical guarantee in Appendix D.3.

4.2 Edge of Stability under Stochastic Training

Even after removing the effect of EoC (i.e., restricting to the positive-curvature subspace in Definition 3.3), the quadratic approximation can still diverge. This divergence is caused by positive curvature when the LR exceeds the critical threshold $\eta_\star(\boldsymbol{w}_{t_0})$. More subtly, neural network training often operates near this boundary: a tiny positive increase in the LR can already trigger divergence.

Stochastic Edge of Stability (EoS) at $\boldsymbol{w}_{t_0}$. Training is at EoS when the current LR is close to the stochastic critical LR on the positive-curvature subspace, $\eta \approx \eta_\star(\boldsymbol{w}_{t_0})$.

When the retained mean curvature is positive definite, the bounded LRs have a single critical endpoint (Appendix D.2). In the full-batch case with a stationary point of the local quadratic, the error obeys $\boldsymbol{x}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{H}_{t_0})\boldsymbol{x}_t$ on the retained subspace. Each positive-curvature component remains bounded exactly when $|1 - \eta\lambda_i| \leq 1$; zero-curvature components remain constant. Hence, if $\lambda_{\max}(\boldsymbol{H}_{t_0}) > 0$, the critical LR is $2/\lambda_{\max}(\boldsymbol{H}_{t_0})$, recovering the full-batch EoS condition of Cohen et al. (2021). Zero modes with affine forcing require separate care, as explained in Appendix D.2. See Appendix A for comparison with previous work on stochastic EoS.

As discussed in Section 3, for the stochastic case, even after projection, mean square stability is governed by the second-moment operator $\mathbb{E}[\boldsymbol{M}_t^+ \otimes \boldsymbol{M}_t^+] = \mathbb{E}[\boldsymbol{M}_t^+] \otimes \mathbb{E}[\boldsymbol{M}_t^+] + \mathrm{Cov}(\boldsymbol{M}_t^+)$. The covariance term can shift the critical LR away from $2/\lambda_{\max}$. Moreover, directly solving $\rho\big(\mathbb{E}[\boldsymbol{M}_t^+ \otimes \boldsymbol{M}_t^+]\big) < 1$ is infeasible at neural network scale, since $\boldsymbol{M}_t^+ \otimes \boldsymbol{M}_t^+$ is a fourth-order object and its entries are affected by noisy mini-batch Hessian samples.

Thus, the definitions above reveal two practical obstacles. First, EoC and EoS live on different curvature subspaces, but explicitly separating these subspaces is infeasible for large NNs, especially for LLM pretraining. Second, the stochastic EoS threshold is not only determined by $\lambda_{\max}$ alone but also correlated with the stochastic noise in training, so we cannot detect it by simply checking whether $\eta\lambda_{\max} \approx 2$.

EoS detection. We use a *local simulation* to detect EoS in pretraining. First, to compensate for negative curvature, we add $\frac{\lambda_P}{2}\|\boldsymbol{w} - \boldsymbol{w}_{t_0}\|^2$ to the local quadratic surrogate (8), with $\lambda_P = \max\{0, -\widehat{\lambda}_{\min}\}$, where $\widehat{\lambda}_{\min}$ is the minimum eigenvalue estimate from Appendix B.2. We then simulate SGD on this loss from the checkpoint for a fixed number of steps over a grid of LRs around the training LR η, using the same batch sequence for every run. The estimate $\widehat{\eta}_\star$ is the smallest LR in the grid for which the displacement norm $\|\boldsymbol{w} - \boldsymbol{w}_{t_0}\|$ reaches a prescribed cutoff. We declare EoS when $\widehat{\eta}_\star$ is close to η. The complete procedure is given in Appendix B.3 and theoretical analysis is in Appendix D.

Illustration on CIFAR-10. Figure 7 applies both detectors to ResNet-20 and ResNet-32 trained with SGD. Each column is a checkpoint, each row an LR factor, and color encodes the final displacement norm of the local simulation. Across most checkpoints in the middle phase of training, the bounded-to-exploding boundary lies at LR factor 1.0 (white dashed line). The training LR is close to $\eta_\star$, and SGD training operates at the EoS for a large part of training. Section 5 applies the same protocol to LLM pretraining with AdamW, with definitions and detectors extended to adaptive methods below.

4.3 Adaptive Gradient Methods at the Edge of Convexity and Edge of Stability

So far, the discussion has focused on SGD. Modern LLMs are usually trained with adaptive methods such as AdamW (Loshchilov & Hutter, 2017). We now consider a general class of adaptive gradient methods (Duchi et al., 2011; Li et al., 2025) as

$$\begin{aligned} \boldsymbol{m}_{t+1} &= \beta_1 \boldsymbol{m}_t + (1-\beta_1)\boldsymbol{g}_{\mathcal{B}_t}(\boldsymbol{w}_t), \\ \boldsymbol{P}_{t+1} &= \Psi(\boldsymbol{P}_t, \boldsymbol{g}_{\mathcal{B}_t}(\boldsymbol{w}_t), \boldsymbol{w}_t), \\ \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \eta \boldsymbol{P}_{t+1}^{-1}\boldsymbol{m}_{t+1} - \eta\lambda_{\mathrm{WD}}\boldsymbol{w}_t, \end{aligned} \tag{AGM}$$

where $0 \le \beta_1 < 1$, $\boldsymbol{m}_t$ is the momentum, $\boldsymbol{P}_t \succ 0$ is the preconditioner matrix, $\Psi(\cdot)$ is the preconditioner update rule, and λ_{WD} is the decoupled weight decay coefficient. AdamW, momentum SGD, and SGD are special cases of this update.

At checkpoint $\boldsymbol{w}_{t_0}$, we freeze $\boldsymbol{P} := \boldsymbol{P}_{t_0}$ and use the local quadratic surrogate (8). The effective curvature and its saddleness are $\boldsymbol{A} := \boldsymbol{P}^{-1/2}\boldsymbol{H}_{t_0}\boldsymbol{P}^{-1/2} + \lambda_{\mathrm{WD}}\boldsymbol{I}, \nu := \max\{0, -\bar{\mu}_{\min}\}/\bar{\mu}_{\max}$, where $\bar{\mu}_{\min} = \lambda_{\min}(\boldsymbol{A})$ and $\bar{\mu}_{\max} = \lambda_{\max}(\boldsymbol{A}) > 0$. Negative effective curvature gives an expanding mean direction for every positive LR, even with momentum (Lemma D.5).

For EoS, we project both momentum and displacement onto the strictly positive eigenspace of $\boldsymbol{A}$ in preconditioned coordinates at every local update. As for SGD, the critical LR is the supremum of LRs for which these projected dynamics are mean square stable. Proposition B.3 shows that this threshold is positive and finite, and gives an equivalent spectral characterization.

Edge of Convexity (EoC) for AGMs. At checkpoint $\boldsymbol{w}_{t_0}$, AGM training is at EoC when $\bar{\mu}_{\min} < 0$ and ν is small.

Edge of Stability (EoS) for AGMs. At checkpoint $\boldsymbol{w}_{t_0}$, AGM training is at EoS when its current LR is close to the unique critical LR of the projected local dynamics, $\eta \approx \eta_\star$.

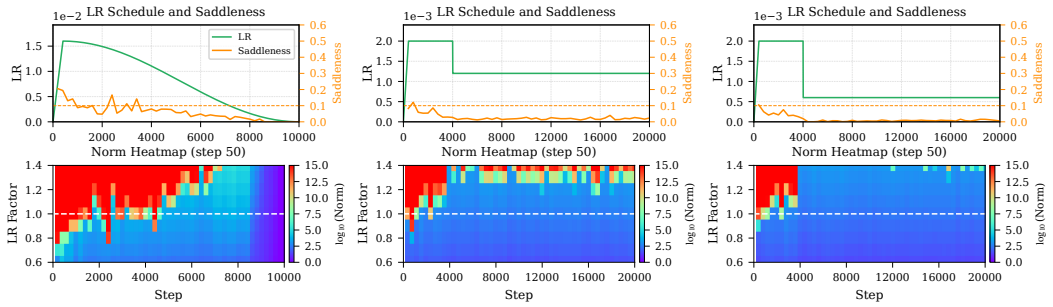
The local model and the definitions of AGM saddleness and critical LR η^* are given in Appendix B.1.

5 Experiments

We apply our detection methods to study EoS and EoC during language model pretraining. Our experiments examine how broadly these phenomena occur and how EoS depends on the training configurations like batch size and learning rate schedule.

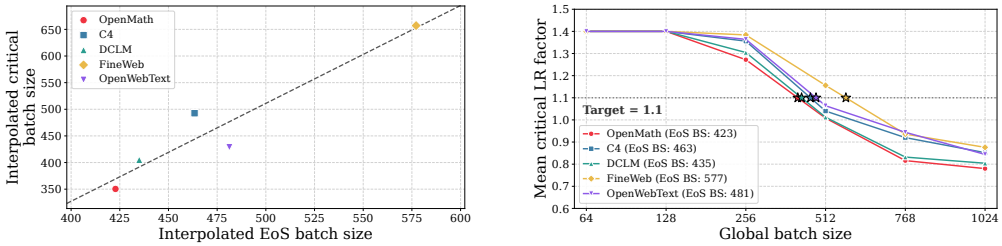
5.1 Experimental Setup

We pretrain GPT-style models at four scales ranging from 43M to 787M. Unless otherwise specified, runs use FineWeb (Penedo et al., 2024), AdamW (WD $= 0.1$, $\beta_1 = 0.9$, $\beta_2 = 0.95$), sequence length 1024, and a Warm–Stable–Decay (WSD) LR schedule (Hu et al., 2024) with 4% warmup steps and 20% decay steps. Full details are provided in Appendix C.



(a) Cosine schedule. (b) Multistep schedule I. (c) Multistep schedule II.

Figure 3: **Training may leave the EoS regime after LR decay.** In (a) and (b), the transition from bounded displacement to explosion moves to larger LR factors during cosine decay and after the abrupt multistep LR drop, respectively. After the larger LR drop in (c), the local simulations remain bounded at the training LR for the rest of the observed run, where training enters the stable regime and does not come back to EoS. See Section 5.4 and Appendices C.4 and C.5 for further discussion.



(a) EoS batch size threshold vs. critical batch size. (b) Critical LR factor q vs. batch size.

Figure 4: **The EoS batch size (EBZ) and the critical batch size (CBZ) show a positive correlation across datasets.** (a) Larger EBZ generally corresponds to larger CBZ, and the dashed line is a linear fit. (b) The critical LR factor $q = \widehat{\eta}_\star/\eta$, averaged over checkpoints, decreases as batch size increases. Stars mark the interpolated crossings of $q = 1.1$, which define EBZ. See Section 5.3 and Appendix C.3 for the experimental protocol and threshold estimation.

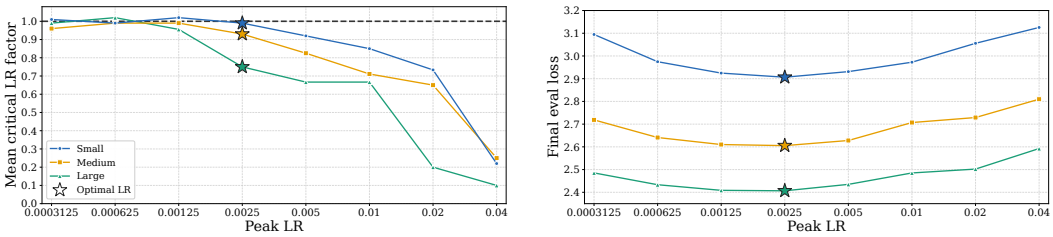
5.2 EoS and EoC Are Widely Observed

We evaluate EoS and EoC at over 3,000 checkpoints from more than 100 pretraining runs spanning different model sizes, peak LRs, batch sizes, and shapes of LR schedules. Both phenomena are widely observed across our experiments (Figures 2, 3, 5, 8 and 9). At many checkpoints, the estimated critical LR is close to the training LR, while the estimated minimum eigenvalue is negative and saddleness is below 0.1, consistent with EoS and EoC, respectively. Additional results and discussion are given in Appendices C.2 to C.5. We next examine how these phenomena depend on the training configuration. Specifically, we focus on two key training hyperparameters: batch size and the LR schedule. Note that EoC remains widely observed as these hyperparameters vary, so the following analysis focuses primarily on how they affect EoS.

5.3 Batch Size Dependence of EoS

For the dependence of batch size, we first show that EoS emerges above a threshold of batch size, which is referred to as the **EoS Batch Size (EBZ)**. We then demonstrate an interesting correlation between EBZ and the critical batch size (CBZ) across different datasets.

EoS emerges above the EBZ threshold. Figure 9 shows that EoS is observed over a larger fraction of training at $B = 512$ and $B = 1024$ than at $B = 256$. For the two larger batch sizes, the transition from bounded displacement to explosion stays near LR factor 1.0 over much of the stable phase, indicating that the training LR is close to the estimated critical LR. Below this batch size threshold, the compensated local simulations at checkpoints after the early phase of training remain bounded at every tested LR factor. These runs use GPT-extra-small on FineWeb with WSD schedules and adjust the LR, weight decay, and β_2 according to the batch size scaling rule in Appendix C.3. Further, across all five datasets, Figure 4b shows that the critical LR factor, averaged over the tested checkpoints of each run, decreases as batch size increases. We estimate EBZ as the batch size at which the fitted



(a) Mean critical LR factor.

(b) Final eval loss.

Figure 5: **Training may go beyond EoS and enter an extremely unstable regime with large LR.** (a) As the peak LR increases, the critical LR factor $q = \widehat{\eta}_\star/\eta$, averaged over the tested checkpoints, falls from near the reference line at 1.0 to well below it. Values near 1.0 indicate EoS, while smaller values mean that the training LR exceeds the estimated critical LR. (b) These large peak LRs also yield higher final evaluation loss. See Sections 5.2 and 5.4 for further discussion.

mean critical LR factor first decreases to 1.1 as batch size increases. This corresponds to the crossing at 1.1 marked in the figure.

EBZ and CBZ show a positive correlation across datasets. In Figure 4a, each point represents one dataset, and the points follow an overall upward linear trend. FineWeb has the largest EBZ and CBZ, while OpenWebMath has the smallest values of both. Here, CBZ is estimated from the same training runs as EBZ and denotes the largest batch size whose final evaluation loss is within 0.3% of the best loss in that dataset's sweep. Thus, datasets with a larger threshold for EoS also tend to maintain near optimal final evaluation loss at larger batch sizes. Full sweep and estimation details are given in Appendix C.3.

5.4 LR Dependence of EoS

We next show that LR decay can cause training to leave EoS, and that training may not return to the EoS regime after a sufficiently large LR decay. We also show that training at excessively large peak LRs can enter an extremely unstable regime beyond EoS.

Training can enter the stable regime after LR decay. Before LR decay, training often operates at EoS, with the training LR close to the estimated critical LR (Figures 3, 11 and 12). After the abrupt LR drop in Figure 3b, the transition from bounded displacement to explosion moves to larger LR factors. The estimated critical LR is now higher than the reduced training LR, and the compensated local simulation stays bounded at the training LR over the tested horizon. As training continues, the estimated critical LR can fall back toward the current training LR and EoS can return, as in Figure 11d. With a sufficiently large LR reduction, however, the estimated critical LR stays well above the training LR for the rest of the observed run (Figures 3c, 11c and 12d). The compensated local simulations then remain bounded at the training LR, and EoS does not return. The effects of the drop magnitude and decay duration are discussed in Appendices C.4 and C.5.

Training can enter an extremely unstable regime with large LR. Training with peak LRs substantially above the empirically optimal range can yield higher loss and enter an *extremely unstable regime* beyond EoS (Figure 5). In this regime, the compensated local quadratic dynamics still allow a critical LR, but it lies far below the training LR. For example, at peak LR 0.04, the estimated critical LR is only about one tenth of the training LR on average, and the compensated local simulations explode at the training LR. See Appendix C.2 for further discussion and Figures 14 to 16 for detailed diagnostics across model sizes and peak LRs.

6 Conclusion

This paper extends the definition of EoS to the LLM training under a unified framework, together with another clearly clarified phenomenon EoC, accounting for the two failure modes of the quadratic approximation. Several directions remain open for future study. (i) Our experiments mainly focus on SGD and AdamW. It is interesting to see how EoS would behave when training with other optimizers like Muon. (ii) Fully understanding EoS and EoC also calls for solid theoretical analyses of the realistic neural network training dynamics.

AI USE STATEMENT

In this work, we used generative AI tools to assist with checking mathematical claims, theoretical derivations, and proofs, and with analyzing and interpreting experimental results. We did not use generative AI tools to design experiments or implement methods. Additionally, we used these tools for manuscript drafting and editing, figure preparation, literature searches, and LaTeX formatting. We reviewed all AI-assisted work, checking the text for accuracy, comparing experimental analyses with the original results, and examining theoretical derivations and proofs for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International conference on machine learning*, pp. 247–257. PMLR, 2022.

Guillaume Alain, Nicolas Le Roux, and Pierre-Antoine Manzagol. Negative eigenvalues of the hessian in deep neural networks. *arXiv preprint arXiv:1902.02366*, 2019.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd. *arXiv preprint arXiv:2412.20553*, 2024.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024. PMLR, 2022.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal*, 19(3):357–367, 1967.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Philippe Bougerol et al. *Products of random matrices with applications to Schrödinger operators*. Springer Science & Business Media, 2012.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Jeremy M Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D Lee. Understanding optimization in deep learning with central flows. *arXiv preprint arXiv:2410.24206*, 2024.

Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

Zhaorui Dong, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

Hillel Furstenberg and Harry Kesten. Products of random matrices. *The Annals of Mathematical Statistics*, 31(2):457–469, 1960. doi: 10.1214/aoms/1177705909. URL https://doi.org/10.1214/aoms/1177705909.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. https://skylion007.github.io/OpenWebTextCorpus/, 2019. URL https://skylion007.github.io/OpenWebTextCorpus/.

Diego Granziol, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size: A random matrix theory approach to neural network training. *Journal of Machine Learning Research*, 23(173):1–65, 2022.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, November 2020. Association for Computational Linguistics.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024.

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. *arXiv preprint arXiv:1807.05031*, 2018.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–50. Springer, 2002.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024. URL https://arxiv.org/abs/2406.11794.

Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold. *arXiv preprint arXiv:2511.02773*, 2025.

Zhouzi Li, Zixuan Wang, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *arXiv preprint arXiv:2207.12678*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *arXiv preprint arXiv:2204.11326*, 2022.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and Scaling Rules for Adaptive Gradient Algorithms, November 2024. URL http://arxiv.org/abs/2205.10287.

Martin Marek, Sanae Lotfi, Aditya Somasundaram, Andrew Gordon Wilson, and Micah Goldblum. Small batch size training for language models: When vanilla sgd works, and why gradient accumulation is wasteful, 2025. URL https://arxiv.org/abs/2507.07101.

Rotem Mulayoff and Tomer Michaeli. Exact mean square linear stability analysis for sgd. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 3915–3969. PMLR, 2024.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06786.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024. URL https://arxiv.org/abs/2406.17557.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, Paris, 1939.

Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36: 74229–74256, 2023.

Lei Wu, Chao Ma, and Weinan E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, 2018a.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018b.

Lei Wu, Mingze Wang, and Weijie Su. The alignment property of sgd noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35:4680–4693, 2022a.

Lei Wu, Mingze Wang, and Weijie J. Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. In *Advances in Neural Information Processing Systems*, 2022b.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pp. 581–590. IEEE, 2020.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need adam: A hessian perspective. *Advances in neural information processing systems*, 37: 131786–131823, 2024.

CONTENTS

1 Introduction **1**

2 Preliminary **2**

3 Stability of Stochastic Gradient Descent on Quadratics **3**

3.1 Negative Curvature Divergence 4

3.2 Positive Curvature Divergence and Critical LR 4

4 Stability of Stochastic Gradient Descent on Neural Networks **5**

4.1 Edge of Convexity 5

4.2 Edge of Stability under Stochastic Training 6

4.3 Adaptive Gradient Methods at the Edge of Convexity and Edge of Stability 7

5 Experiments **7**

5.1 Experimental Setup 7

5.2 EoS and EoC Are Widely Observed 8

5.3 Batch size dependence of EoS 8

5.4 LR Dependence of EoS 9

6 Conclusion **9**

A Related Work **15**

B Detailed Methods **16**

B.1 Adaptive Gradient Methods 16

B.2 EoC Detection 18

B.3 EoS Detection 21

B.4 Cost of a Hessian-Vector Product 22

C Experimental Details **23**

C.1 Models, Training, and Detection Settings 23

C.2 Additional Results 24

C.3 Batch Size Threshold Experiments 25

C.4 Multistep LR Drop Ablations 25

C.5 Linear LR Decay Ablations 25

C.6 Experiments Across Model Scales 25

D Theoretical Analysis **34**

D.1 Setting and the Local Linear System 34

D.2 Stability of the Local Dynamics 36

D.3 Stochastic Power Iteration for EoC 38
D.4 Compensated Local Simulation for EoS 41

E Proofs for the Local Linear System and the Stability Results **44**
E.1 Proof of Proposition D.2 44
E.2 Proof of Proposition D.3 45
E.3 Proof of Corollary D.4 46
E.4 Proofs of Lemma D.5 and Corollary D.6 46
E.5 Proof of Corollary D.8 47
E.6 Proof of the AGM Critical LR Characterization 49

F Proof of the Power-Iteration Bound **50**
F.1 The Bisection Bound 50
F.2 Proof of the Explosion Bound 51
F.3 Proof of the No-Explosion Bound 53
F.4 Auxiliary Results 55

G Proof of the Compensated Finite-Horizon Threshold Bound **56**
G.1 Stability Criterion 56
G.2 Proof of the Threshold Bound 59
G.3 Compensation Error 60
G.4 Finite-Horizon Error 70
G.4.1 Bounding the Moment Radius 70
G.4.2 Bounding the Inverse LR 73
G.4.3 Choosing the Cutoff 75
G.5 Proofs of the Stability Criterion 77
G.5.1 Proof of the Exact Criterion 77
G.5.2 Proof of the Quantitative Bound 82
G.5.3 Proof of the AGM Critical LR 86
G.6 Auxiliary Results 87

A Related Work

Edge of Stability. Early studies connected sharpness and stability to optimizer behavior in neural network training: Wu et al. (2018b) analyzed SGD minima selection through dynamical stability using sharpness and non-uniformity, and Xing et al. (2018) observed SGD "bouncing" between valley walls during training. Jastrzębski et al. (2018) then showed that SGD visits increasingly sharp regions up to a maximum set by LR and batch size, and that the step is often too large along the sharpest directions. Cohen et al. (2021) later demonstrated that full-batch gradient descent typically operates at an Edge of Stability (EoS), where the largest Hessian eigenvalue oscillates near $2/\eta$ while loss decreases non-monotonically. Cohen et al. (2022) extended this phenomenon to adaptive methods via saturation of a preconditioned stability threshold. And theoretical explanations for EoS were discussed in simplified models or realistic assumptions(Arora et al., 2022; Li et al., 2022; Damian et al., 2022).

Training Stability of Neural Networks Series works studied the training stability of constant LR SGD in quadratic losses via a linear system analysis of the second moment of the parameters (Wu et al., 2018b; Ma & Ying, 2021; Granziol et al., 2022; Wu et al., 2022a; Mulayoff & Michaeli, 2024). They obtained tight conditions in the form of $\|\mathbb{E}[(\boldsymbol{I} - \eta \boldsymbol{H}_t)^{\otimes 2}]\| \leq 1$, which are valid instability criteria for a particular Lyapunov function. However, in the training neural networks with (S)GD, the unstable convergence phenomenon was observed (Wu et al., 2018b; Xing et al., 2018; Jastrzębski et al., 2018), formally identified and termed *Edge of Stability* (EoS) (Cohen et al., 2021). Subsequently, several studies sought to explain this unstable convergence (Ma et al., 2022; Ahn et al., 2022; Wu et al., 2023; Arora et al., 2022; Damian et al., 2022; Cohen et al., 2024). Among them, Cohen et al. (2024) gave a general characterization of the oscillatory dynamics of deep learning in the EoS regime via a time-averaged differential equation termed *central flow*. However, the effectiveness of the quadratic approximation was never carefully studied in the large-scale neural network training, hence it is not clear when and how the quadratic approximation fails or not in the realistic regime.

Hessian Spectrum of Neural Networks. Most studies on the Hessian spectrum of neural networks reported that the Hessian spectra of neural networks consist of a "bulk" together with a few large "outliers" in magnitude (LeCun et al., 2002; Dauphin et al., 2014; Sagun et al., 2016; 2017; Ghorbani et al., 2019; Chaudhari et al., 2019; Yao et al., 2020). Among them, Sagun et al. (2016; 2017) mentioned that there are still negative eigenvalues during training, but their magnitude is much smaller than the large outliers. More recent works reported and explained a near-block diagonal structure in the Hessian spectrum of CNNs and Transformers (Zhang et al., 2024; Dong et al., 2025). However, one thing that has long been overlooked is the characterization of the negative eigenvalues in the population Hessian during large-scale training. In this work, we present detailed empirical findings of the negative eigenvalues across different setups in large-scale neural network training.

Comparison with Andreyev & Beneventano (2024). The recent work (Andreyev & Beneventano, 2024) is among the most relevant to this work, which defines Edge of Stochastic Stability for SGD via the batch sharpness, where Batch Sharpness hovers around $2/\eta$. Compared to Andreyev & Beneventano (2024), our work (1) In the full-batch setting, our definition for EoS can be reduced to the original definition of EoS by Cohen et al. (2021; 2022), while theirs cannot; (2) We also identify another failure mode of the quadratic approximation called Edge of Convexity (EoC); (3) We provide scalable EoS and EoC detection methods with theoretical guarantees; (4) We conduct experiments on large-scale LLM pretraining with AdamW, while most of their experiments are limited to ResNets on CIFAR-10 and SVHN with vanilla SGD.

B Detailed Methods

We first define the local AGM model, its saddleness, and its critical LR in Appendix B.1. Appendices B.2 and B.3 then describe the EoC and EoS detection procedures and their guarantees. Both procedures use Hessian-vector products without forming the Hessian. Appendix B.4 gives their computational cost. The complete theoretical statements and proofs are in Appendices D to G.

B.1 Adaptive Gradient Methods

We consider adaptive gradient methods (AGMs) with a momentum vector $\boldsymbol{m}_t$ and a positive definite preconditioner $\boldsymbol{P}_t$. For the mini-batch loss $\mathcal{L}_{\mathcal{B}_t}$, the updates are

$$\begin{aligned}\boldsymbol{m}_{t+1} &= \beta_1 \boldsymbol{m}_t + (1-\beta_1)\nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_t),\\ \boldsymbol{P}_{t+1} &= \Psi\big(\boldsymbol{P}_t, \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_t), \boldsymbol{w}_t\big),\\ \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \eta \boldsymbol{P}_{t+1}^{-1}\boldsymbol{m}_{t+1} - \eta\lambda_{\mathrm{WD}}\boldsymbol{w}_t.\end{aligned}$$

Here $\eta > 0$ is the LR, $0 \le \beta_1 < 1$ is the momentum coefficient, and $\lambda_{\mathrm{WD}} \ge 0$ is the decoupled weight decay coefficient. The preconditioner update Ψ preserves positive definiteness.

The local quadratic model. Fix a checkpoint $\boldsymbol{w}_{t_0}$. For each mini-batch $\mathcal{B}_t$, we approximate its loss $\mathcal{L}_{\mathcal{B}_t}$ by the second-order Taylor expansion around this checkpoint. Write $\boldsymbol{\epsilon} = \boldsymbol{w} - \boldsymbol{w}_{t_0}$ for the displacement, and define the gradient and Hessian at the checkpoint by

$$\boldsymbol{g}_{\mathcal{B}_t} \coloneqq \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}), \qquad \boldsymbol{H}_{\mathcal{B}_t} \coloneqq \nabla^2 \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}).$$

The resulting *quadratic surrogate* is

$$Q_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}+\boldsymbol{\epsilon}) \coloneqq \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) + \boldsymbol{g}_{\mathcal{B}_t}^\top \boldsymbol{\epsilon} + \tfrac{1}{2}\boldsymbol{\epsilon}^\top \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon}. \tag{11}$$

This surrogate matches the loss, gradient, and Hessian at the checkpoint and omits terms of higher order in the displacement. At each local step, we sample a fresh batch and use its surrogate in place of the neural network loss. The coefficients $\boldsymbol{g}_{\mathcal{B}_t}$ and $\boldsymbol{H}_{\mathcal{B}_t}$ are always evaluated at $\boldsymbol{w}_{t_0}$.

Local AGM dynamics. We freeze the preconditioner at $\boldsymbol{P} \coloneqq \boldsymbol{P}_{t_0} \succ 0$ and let the momentum evolve under the surrogate gradients. Let $\boldsymbol{\epsilon}_t$ be the displacement after t local steps. The surrogate gradient at this displacement is $\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon}_t$. Keeping η, β_1 and λ_{WD} fixed, the local updates become

$$\begin{aligned}\boldsymbol{m}_{t+1} &= \beta_1 \boldsymbol{m}_t + (1-\beta_1)(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon}_t),\\ \boldsymbol{\epsilon}_{t+1} &= \boldsymbol{\epsilon}_t - \eta \boldsymbol{P}^{-1}\boldsymbol{m}_{t+1} - \eta\lambda_{\mathrm{WD}}(\boldsymbol{w}_{t_0}+\boldsymbol{\epsilon}_t).\end{aligned} \tag{12}$$

The simulation starts at $\boldsymbol{\epsilon}_0 = \boldsymbol{0}$ with the checkpoint momentum. Weight decay acts on the full parameter $\boldsymbol{w}_{t_0} + \boldsymbol{\epsilon}_t$. The choice $\beta_1 = 0$, $\boldsymbol{P} = \boldsymbol{I}$, and $\lambda_{\mathrm{WD}} = 0$ recovers SGD.

In preconditioned coordinates, we stack the momentum and displacement as

$$\boldsymbol{z}_t \coloneqq (\eta \boldsymbol{P}^{-1/2}\boldsymbol{m}_t, \boldsymbol{P}^{1/2}\boldsymbol{\epsilon}_t)^\top.$$

Then (12) has the affine form

$$\boldsymbol{z}_{t+1} = \boldsymbol{M}_t(\eta)\boldsymbol{z}_t + \boldsymbol{q}_t. \tag{13}$$

Write $\boldsymbol{D}_{\mathcal{B}_t} = \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2}$ for the preconditioned mini-batch Hessian. The transition matrix and forcing are

$$\boldsymbol{M}_t(\eta) = \begin{bmatrix} \beta_1 \boldsymbol{I} & (1-\beta_1)\eta \boldsymbol{D}_{\mathcal{B}_t} \\ -\beta_1 \boldsymbol{I} & (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{I} - (1-\beta_1)\eta \boldsymbol{D}_{\mathcal{B}_t} \end{bmatrix},$$

$$\boldsymbol{q}_t = \begin{bmatrix} (1-\beta_1)\eta \boldsymbol{P}^{-1/2}\boldsymbol{g}_{\mathcal{B}_t} \\ -(1-\beta_1)\eta \boldsymbol{P}^{-1/2}\boldsymbol{g}_{\mathcal{B}_t} - \eta\lambda_{\mathrm{WD}}\boldsymbol{P}^{1/2}\boldsymbol{w}_{t_0} \end{bmatrix}.$$

Two trajectories driven by the same batches have the same forcing $\boldsymbol{q}_t$. Their difference therefore evolves by multiplication with $\boldsymbol{M}_t(\eta)$. We use this homogeneous evolution to define stability of local perturbations.

Effective curvature and saddleness. Let $\boldsymbol{H}_{t_0} \coloneqq \nabla^2 \mathcal{L}(\boldsymbol{w}_{t_0})$ be the population Hessian. The preconditioned Hessians and effective curvatures are

$$\begin{aligned} \boldsymbol{D}_{t_0} &\coloneqq \boldsymbol{P}^{-1/2} \boldsymbol{H}_{t_0} \boldsymbol{P}^{-1/2}, \qquad \boldsymbol{A} \coloneqq \boldsymbol{D}_{t_0} + \lambda_{\mathrm{WD}} \boldsymbol{I}, \\ \boldsymbol{D}_{\mathcal{B}_t} &\coloneqq \boldsymbol{P}^{-1/2} \boldsymbol{H}_{\mathcal{B}_t} \boldsymbol{P}^{-1/2}, \quad \boldsymbol{A}_{\mathcal{B}_t} \coloneqq \boldsymbol{D}_{\mathcal{B}_t} + \lambda_{\mathrm{WD}} \boldsymbol{I}. \end{aligned} \tag{14}$$

If $\lambda_{\min}(\boldsymbol{A}) < 0$, some local perturbations grow in mean at every positive LR, even with momentum (Lemma D.5). We measure the size of this negative effective curvature relative to the largest positive effective curvature.

Definition B.1 (Saddleness for AGMs)**.** Suppose $\lambda_{\max}(\boldsymbol{A}) > 0$. The AGM saddleness at the checkpoint is

$$\nu \coloneqq \frac{\max\{0, -\lambda_{\min}(\boldsymbol{A})\}}{\lambda_{\max}(\boldsymbol{A})}. \tag{15}$$

The preconditioner and weight decay are included through $\boldsymbol{A}$. AGM training is at EoC when $\lambda_{\min}(\boldsymbol{A}) < 0$ and this ratio is small. For SGD without weight decay, $\boldsymbol{A} = \boldsymbol{H}_{t_0}$, so this definition reduces to Definition 4.1.

Critical LR. To isolate instability caused by the LR, we remove the directions of nonpositive effective curvature. Let Π^+ project onto the strictly positive eigenspace of $\boldsymbol{A}$, assumed to be nonempty. We project both momentum and displacement in preconditioned coordinates at every update. The retained effective curvature is therefore positive definite. Write $\boldsymbol{M}_t^+(\eta)$ and $\boldsymbol{q}_t^+$ for the transition and forcing in the retained coordinates. The projected local system is

$$\boldsymbol{z}_{t+1}^+ = \boldsymbol{M}_t^+(\eta) \boldsymbol{z}_t^+ + \boldsymbol{q}_t^+. \tag{16}$$

Following Definition 3.1, we call this system mean square stable if

$$\sup_{t \geq 0} \mathbb{E}\|\boldsymbol{z}_t^+\|^2 < \infty \qquad \text{for every deterministic initial state } \boldsymbol{z}_0^+.$$

The initial state includes both momentum and displacement. This definition allows persistent fluctuations and does not require the second moment to tend to zero.

Definition B.2 (Critical LR for AGMs)**.** Keep the checkpoint, $\boldsymbol{P}$, β_1, λ_{WD}, and the batch distribution fixed as the LR varies. The AGM critical LR is

$$\eta_\star \coloneqq \sup\big\{\eta > 0 : \text{the projected system (16) is mean square stable}\big\}. \tag{17}$$

The next proposition gives an equivalent way to determine this critical LR from the homogeneous transition. Its second-moment matrix is

$$\boldsymbol{K}^+(\eta) \coloneqq \mathbb{E}[\boldsymbol{M}_t^+(\eta) \otimes \boldsymbol{M}_t^+(\eta)]. \tag{18}$$

We assume that the pairs $(\boldsymbol{H}_{\mathcal{B}_t}, \boldsymbol{g}_{\mathcal{B}_t})$ are i.i.d. and independent of the initial state. Both entries have finite second moments, each Hessian is symmetric, and their means are the population derivatives at the checkpoint. These are the mini-batch assumptions of Assumption D.1.

Proposition B.3 (Spectral characterization of the AGM critical LR)**.** *Under Assumption D.1, fix $\boldsymbol{P} \succ 0$, $0 \leq \beta_1 < 1$ and $\lambda_{\mathrm{WD}} \geq 0$. Suppose the strictly positive eigenspace of $\boldsymbol{A}$ is nonempty. The critical LR in Definition B.2 satisfies*

$$0 < \eta_\star = \sup\big\{\eta > 0 : \rho(\boldsymbol{K}^+(\eta)) < 1\big\} < \infty. \tag{19}$$

Moreover,

$$\rho\big(\boldsymbol{K}^+(\eta)\big) \begin{cases} < 1, & 0 < \eta < \eta_\star, \\ = 1, & \eta = \eta_\star, \\ > 1, & \eta > \eta_\star. \end{cases} \tag{20}$$

Below $\eta_\star$, the projected affine system is mean square stable. Above $\eta_\star$, some deterministic initialization has an unbounded second moment. At $\eta_\star$, boundedness can depend on the forcing. Thus the proposition equates the two critical LRs without identifying boundedness with strict decay at the boundary. The proof is in Appendix E.6. AGM training is at EoS when its training LR is close to $\eta_\star$.

The stability notion is the same as in the main text, but the AGM projection retains only strictly positive effective curvature. The SGD projection in Definition 3.3 also retains zero curvature. For SGD without weight decay, the two critical LRs agree when the population Hessian has no zero eigenvalue. Appendix D.2 explains the distinction when zero modes remain. The next two subsections describe how we estimate the effective curvature and the critical LR of this local model.

B.2 EoC Detection

EoC at $\boldsymbol{w}_{t_0}$ means that $\boldsymbol{A}$ has a negative eigenvalue while the saddleness ν is small. Detecting it requires $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$. We estimate the maximum eigenvalue by Lanczos on a large batch. In our measurements, the top eigenpair varies little across mini-batches, whereas the bottom eigenvectors are nearly orthogonal (Figure 6). Lanczos on one mini-batch estimates $\lambda_{\min}(\boldsymbol{A}_{\mathcal{B}_t})$, whose expectation is at most $\lambda_{\min}(\boldsymbol{A})$ by Jensen's inequality. We therefore estimate $\lambda_{\min}(\boldsymbol{A})$ by an iteration that draws a fresh mini-batch at every step.

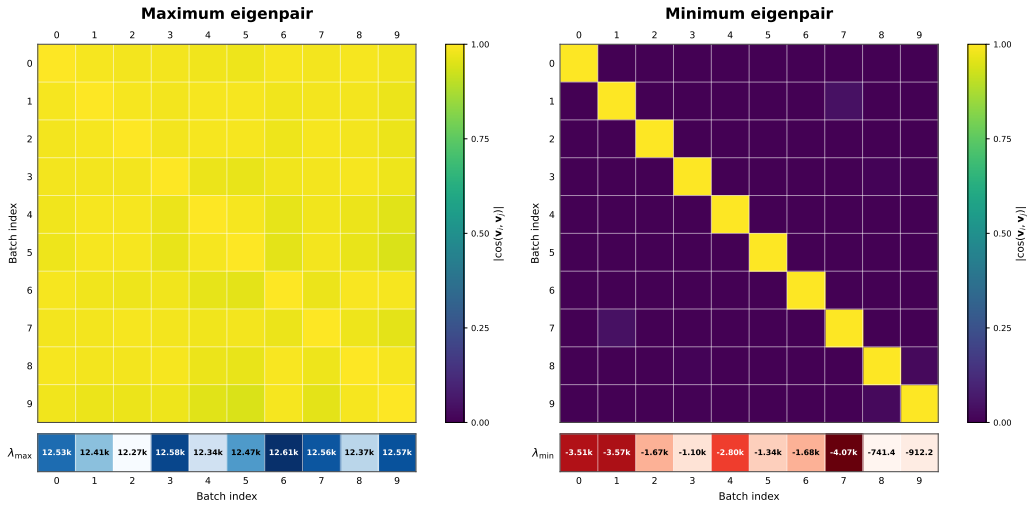


Figure 6: **Mini-batch Lanczos estimates are stable for the top eigenpair but unstable for the bottom eigenpair.** We run Lanczos on ten independently sampled mini-batches at the same checkpoint for the Adam-preconditioned Hessian. The maximum eigenvectors have high pairwise absolute cosine and similar eigenvalues across batches. In contrast, the minimum eigenvectors are nearly orthogonal across batches and the corresponding $\lambda_{\min}$ values vary substantially, showing that per-batch Lanczos estimates are dominated by batch-specific negative-curvature directions.

Stochastic power iteration with bisection. Algorithm 1 runs the shifted iteration on the mini-batch effective curvatures,

$$\boldsymbol{x}_{t+1} = \big(\boldsymbol{I} - \eta_{\mathrm{p}}(\boldsymbol{A}_{\mathcal{B}_t} + \lambda \boldsymbol{I})\big)\boldsymbol{x}_t, \qquad t = 0, \dots, T_{\mathrm{p}} - 1,$$

with a fresh mini-batch at every step, each $\boldsymbol{A}_{\mathcal{B}_t}\boldsymbol{x}_t$ formed from one Hessian-vector product. Since the batches are independent, the mean iterate follows the population matrix, $\mathbb{E}[\boldsymbol{x}_{t+1}] = (\boldsymbol{I} - \eta_{\mathrm{p}}(\boldsymbol{A} + \lambda \boldsymbol{I}))\mathbb{E}[\boldsymbol{x}_t]$. If $\lambda_{\min}(\boldsymbol{A}) + \lambda < 0$, this matrix has an eigenvalue larger than one. When the initial vector has nonzero overlap with a bottom eigenvector, the component of $\mathbb{E}[\boldsymbol{x}_t]$ along that eigenvector grows geometrically. If $\lambda_{\min}(\boldsymbol{A}) + \lambda \geq 0$ and η_{p} is small enough, $\mathbb{E}[\boldsymbol{x}_t]$ stays bounded. A query at shift λ is therefore declared to explode when $\|\boldsymbol{x}_t\|$ exceeds the cutoff B at some step $t \leq T_{\mathrm{p}}$, a relative cutoff since $\|\boldsymbol{x}_0\| = 1$. Bisection on a supplied bracket $[0, \Lambda]$ moves the lower endpoint to an exploding shift and the upper endpoint to a non-exploding shift. After N rounds the output $\widehat{\lambda}$ is the final upper endpoint, and the minimum-eigenvalue estimate is $\widehat{\lambda}_{\min} = -\widehat{\lambda}$.

Three choices matter in practice: the initial vector, the stepsize, and the bracket. The initial vector $\boldsymbol{x}_0$ is the same unit vector for every query. For nonzero checkpoint momentum, we use $\boldsymbol{m}_{t_0}/\|\boldsymbol{m}_{t_0}\|$. The stepsize η_{p} must satisfy $\eta_{\mathrm{p}}(\lambda_{\max}(\boldsymbol{A}) + \Lambda) \leq 1$ over the whole bracket. This makes the mean transition positive semidefinite for every queried shift and keeps $\|\boldsymbol{I} - \eta_{\mathrm{p}}(\boldsymbol{A} + \lambda \boldsymbol{I})\| \leq 1$ for every $\lambda \in [\lambda_\star, \Lambda]$, where $\lambda_\star := \max\{0, -\lambda_{\min}(\boldsymbol{A})\}$. We set it from the Lanczos estimate $\widehat{\lambda}_{\max}$, which Algorithm 2 computes first on $\bar{\boldsymbol{A}}$, the effective curvature averaged over a fixed set of mini-batches. This estimate does not certify the population stepsize condition. The bracket $[0, \Lambda]$ is supplied to the algorithm and must contain $\lambda_\star$. The guarantee takes it as given, assumes $\lambda_{\min}(\boldsymbol{A}) < 0$, and includes the bisection resolution $\Lambda 2^{-N}$. We use $B \geq 2$. The algorithm also runs when $\lambda_{\min}(\boldsymbol{A}) \geq 0$, but the guarantee below does not cover that case. Each query uses at most T_{p} Hessian-vector products.

Algorithm 1 Minimum eigenvalue estimation by stochastic power iteration and bisection

Require: Symmetric samples $\boldsymbol{A}_{\mathcal{B}_t}$, i.i.d. with mean $\boldsymbol{A}$; inner stepsize η_{p}; horizon T_{p}; bisection rounds N; cutoff $B \geq 2$; a supplied unit vector $\boldsymbol{x}_0$; and a bracket $[0, \Lambda]$ containing $\lambda_\star = \max\{0, -\lambda_{\min}(\boldsymbol{A})\}$ with $\eta_{\mathrm{p}}(\lambda_{\max}(\boldsymbol{A}) + \Lambda) \leq 1$.

Ensure: Estimated shift $\widehat{\lambda}$ and minimum eigenvalue estimate $\widehat{\lambda}_{\min} = -\widehat{\lambda}$.

Subroutine STABLE(λ):
    Set $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$; use fresh independent samples for this query.
for $t = 0, 1, \ldots, T_{\mathrm{p}} - 1$ **do**
    Draw a fresh mini-batch $\mathcal{B}_t$ and form $\boldsymbol{A}_{\mathcal{B}_t}$.
    $\boldsymbol{x} \leftarrow \boldsymbol{x} - \eta_{\mathrm{p}}(\boldsymbol{A}_{\mathcal{B}_t} + \lambda \boldsymbol{I})\boldsymbol{x}$.
    if $\|\boldsymbol{x}\| > B$ **then**
        return `false` {explode}
    end if
end for
return `true` {stable}

Main procedure:
$\lambda_{\mathrm{lo}} \leftarrow 0, \quad \lambda_{\mathrm{hi}} \leftarrow \Lambda$. {use the supplied bracket}
for $k = 1, 2, \ldots, N$ **do**
    $\lambda_{\mathrm{mid}} \leftarrow (\lambda_{\mathrm{lo}} + \lambda_{\mathrm{hi}})/2$.
    if STABLE(λ_{mid}) = `true` **then**
        $\lambda_{\mathrm{hi}} \leftarrow \lambda_{\mathrm{mid}}$ {stable}
    else
        $\lambda_{\mathrm{lo}} \leftarrow \lambda_{\mathrm{mid}}$ {explode}
    end if
end for
$\widehat{\lambda} \leftarrow \lambda_{\mathrm{hi}}, \quad \widehat{\lambda}_{\min} \leftarrow -\widehat{\lambda}$.
return $\widehat{\lambda}$, $\widehat{\lambda}_{\min}$.

Algorithm 2 EoC detection for AGMs

Require: Checkpoint $\boldsymbol{w}_{t_0}$; frozen preconditioner $\boldsymbol{P} = \boldsymbol{P}_{t_0} \succ 0$; weight decay $\lambda_{\mathrm{WD}} \geq 0$; tolerance $\nu_{\mathrm{tol}} > 0$; momentum state $\boldsymbol{m}_{t_0}$; Algorithm 1 parameters $(\eta_{\mathrm{p}}, T_{\mathrm{p}}, N, B, \Lambda)$; Lanczos iterations T_{LZ}; averaged operator $\bar{\boldsymbol{A}}$, the mean of $\boldsymbol{A}_{\mathcal{B}}$ over a fixed set of mini-batches.

Ensure: EoC $\in \{\mathtt{true}, \mathtt{false}\}$; estimates $\widehat{\lambda}_{\min}$, $\widehat{\lambda}_{\max}$, and $\widehat{\nu}$.

Define symmetric stochastic samples $\boldsymbol{A}_{\mathcal{B}_t} := \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2} + \lambda_{\mathrm{WD}}\boldsymbol{I}$.
$\widehat{\lambda}_{\max} \leftarrow$ LANCZOSMAXEIG($\bar{\boldsymbol{A}}$; T_{LZ}). {η_{p} is chosen from $\widehat{\lambda}_{\max}$, see the text}
Set $\boldsymbol{x}_0 \leftarrow \boldsymbol{m}_{t_0}/\|\boldsymbol{m}_{t_0}\|$.
$(\widehat{\lambda}, \widehat{\lambda}_{\min}) \leftarrow$ Algorithm 1$(\boldsymbol{A}_{\mathcal{B}_t}; \eta_{\mathrm{p}}, T_{\mathrm{p}}, N, B, \Lambda, \boldsymbol{x}_0)$.
$\widehat{\nu} \leftarrow \dfrac{\max\{0, -\widehat{\lambda}_{\min}\}}{\widehat{\lambda}_{\max}}$.
EoC $\leftarrow \big(\widehat{\lambda}_{\min} < 0\big) \wedge \big(\widehat{\nu} \leq \nu_{\mathrm{tol}}\big)$.
return EoC, $\widehat{\lambda}_{\min}$, $\widehat{\lambda}_{\max}$, $\widehat{\nu}$.

Algorithms 2 and 3 pass T_{p} to Algorithm 1, and Algorithm 3 writes T for its own simulation horizon. The horizons used in the experiments are listed in Appendix C.1.

EoC criterion. Algorithm 2 combines the two eigenvalue estimates into the saddleness estimate $\widehat{\nu} = \max\{0, -\widehat{\lambda}_{\min}\}/\widehat{\lambda}_{\max}$, which estimates the AGM saddleness ν of Definition B.1. We declare EoC when $\widehat{\lambda}_{\min} < 0$ and $\widehat{\nu} \leq \nu_{\mathrm{tol}}$. This ratio requires $\widehat{\lambda}_{\max} > 0$. The experiments use $\nu_{\mathrm{tol}} = 0.1$. The same estimate sets the compensation coefficient $\lambda_{\mathrm{P}} = \max\{0, -\widehat{\lambda}_{\min}\}$ for EoS detection.

Guarantee. The following theorem separates the error due to curvature noise, the finite horizon, and the bisection resolution.

Theorem B.4 (Informal; accuracy of the stochastic power iteration)**.** *Let $\lambda_{\min}(\boldsymbol{A}) < 0$, and let the mini-batch effective curvatures $\boldsymbol{A}_{\mathcal{B}_t} = \boldsymbol{A} + \boldsymbol{\Delta}_t$ be i.i.d. symmetric with $\mathbb{E}[\boldsymbol{\Delta}_t] = \mathbf{0}$ and $\|\boldsymbol{\Delta}_t\| \le \sigma$ almost surely. Let the unit initialization $\boldsymbol{x}_0$ satisfy $|\boldsymbol{u}_{\min}^\top \boldsymbol{x}_0| \ge \varepsilon > 0$ for a unit eigenvector $\boldsymbol{u}_{\min}$ of $\boldsymbol{A}$ with eigenvalue $\lambda_{\min}(\boldsymbol{A})$, and let the bracket and stepsize conditions of Theorem D.12 hold. For fixed confidence $1-\delta$, Algorithm 1 with horizon T_{p}, cutoff B and N bisection rounds outputs, with probability at least $1-\delta$,*

$$\left|\widehat{\lambda}_{\min} - \lambda_{\min}(\boldsymbol{A})\right| \le O\big(\eta_{\mathrm{p}} \log(2N)\big) + O\Big(\frac{1}{\eta_{\mathrm{p}} T_{\mathrm{p}}}\Big) + \Lambda 2^{-N},$$

where the hidden constants depend on the noise bound σ, cutoff B, initial overlap ε, and failure probability δ.

The first term comes from curvature noise, the second from running each query for only T_{p} steps, and the third from the number of bisection rounds. For large T_{p}, choosing η_{p} of order $[T_{\mathrm{p}} \log(2N)]^{-1/2}$ balances the first two terms and gives $O(\sqrt{\log(2N)/T_{\mathrm{p}}})$, plus the bisection error. Increasing the horizon at a fixed stepsize reduces the finite-horizon term but does not remove the noise term from this bound. The theorem does not establish the bracket or the overlap. Both are inputs, and the theorem is silent when $\lambda_{\min}(\boldsymbol{A}) \ge 0$. The formal statement is Theorem D.12 in Appendix D.3, and the proof is in Appendix F.

Algorithm 3 EoS detection for AGMs

Require: Checkpoint $\boldsymbol{w}_{t_0}$; momentum state $\boldsymbol{m}_{t_0}$; frozen preconditioner $\boldsymbol{P} = \boldsymbol{P}_{t_0} \succ 0$; weight decay $\lambda_{\mathrm{WD}} \geq 0$; training LR η; momentum parameter $\beta_1 \in [0, 1)$; horizon T; cutoff M; random seed s; increasing LR factors $\{\alpha_k\}_{k=1}^{K}$; EoS tolerance $\tau_{\mathrm{eos}} > 0$; parameters $(\eta_{\mathrm{p}}, T_{\mathrm{p}}, N, B, \Lambda, \boldsymbol{x}_0)$ for Algorithm 1, with $\boldsymbol{x}_0 = \boldsymbol{m}_{t_0}/\|\boldsymbol{m}_{t_0}\|$.

Ensure: Estimated critical LR $\widehat{\eta}_\star$, critical LR ratio $q(\boldsymbol{w}_{t_0}) = \widehat{\eta}_\star/\eta$, and EoS flag.

(Step 1: estimate the compensation coefficient.) Define symmetric stochastic samples $\boldsymbol{A}_{\mathcal{B}_t} := \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2} + \lambda_{\mathrm{WD}}\boldsymbol{I}$. Run Algorithm 1 on $\boldsymbol{A}_{\mathcal{B}_t}$ to obtain $\widehat{\lambda}_{\min}$. Set $\lambda_{\mathsf{P}} \leftarrow \max\{0, -\widehat{\lambda}_{\min}\}$.

Subroutine ExplodeDetect(α):

  Set the random seed to s; $\boldsymbol{\epsilon} \leftarrow \boldsymbol{0}$; $\boldsymbol{m} \leftarrow \boldsymbol{m}_{t_0}$; $R \leftarrow 0$.

for $t = 0, 1, \ldots, T-1$ **do**

  Sample mini-batch $\mathcal{B}_t$.

  $\boldsymbol{g} \leftarrow \boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon} + \lambda_{\mathsf{P}}\boldsymbol{P}\boldsymbol{\epsilon}$.

  $\boldsymbol{m} \leftarrow \beta_1\boldsymbol{m} + (1-\beta_1)\boldsymbol{g}$.

  $\boldsymbol{\epsilon} \leftarrow \boldsymbol{\epsilon} - \alpha\eta\boldsymbol{P}^{-1}\boldsymbol{m} - \alpha\eta\lambda_{\mathrm{WD}}(\boldsymbol{\epsilon} + \boldsymbol{w}_{t_0})$.

  $R \leftarrow \max\{R, \|\boldsymbol{\epsilon}\|_{\boldsymbol{P}}\}$.

  if $R \geq M$ **then**

    return `true` {explode}

  end if

end for

  return `false` {stable}

(Step 2: search over $\{\alpha_k\}$.) Find $k^\star := \min\{k : \textsc{ExplodeDetect}(\alpha_k) = \texttt{true}\}$.

if no such $k^\star$ exists **then**

  Record the censored bounds $\widehat{\eta}_\star > \alpha_K\eta$ and $q > \alpha_K$; set EoS $\leftarrow$ `false`.

  return $\widehat{\eta}_\star, q,$ EoS.

end if

$\widehat{\eta}_\star \leftarrow \alpha_{k^\star}\eta$.

$q(\boldsymbol{w}_{t_0}) \leftarrow \alpha_{k^\star}$.

EoS $\leftarrow \big(q(\boldsymbol{w}_{t_0}) \in [1 - \tau_{\mathrm{eos}},\ 1 + \tau_{\mathrm{eos}}]\big)$.

return $\widehat{\eta}_\star, q,$ EoS.

B.3 EoS Detection

The AGM critical LR is the threshold for bounded second moments on the strictly positive eigenspace of $\boldsymbol{A}$ (Definition B.2). Forming this projection is impractical at LLM scale, and running the local dynamics without it is misleading, because negative effective curvature produces an expanding direction at every positive LR (Lemma D.5). We therefore add compensation to the local quadratic loss and locate the stability boundary by simulation.

Compensated local loss. Let $\boldsymbol{\epsilon} := \boldsymbol{w} - \boldsymbol{w}_{t_0}$ be the displacement from the checkpoint. With λ_{P} from Appendix B.2, the compensated local loss is the quadratic surrogate (11) with the penalty measured in the norm of the frozen preconditioner,

$$G_t(\boldsymbol{\epsilon}) := \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) + \boldsymbol{\epsilon}^\top\boldsymbol{g}_{\mathcal{B}_t} + \tfrac{1}{2}\boldsymbol{\epsilon}^\top\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon} + \tfrac{\lambda_{\mathsf{P}}}{2}\|\boldsymbol{\epsilon}\|_{\boldsymbol{P}}^2, \qquad \|\boldsymbol{\epsilon}\|_{\boldsymbol{P}}^2 = \boldsymbol{\epsilon}^\top\boldsymbol{P}\boldsymbol{\epsilon}.$$

The penalty shifts the mini-batch effective curvature from $\boldsymbol{A}_{\mathcal{B}_t}$ to $\boldsymbol{A}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I}$ in preconditioned coordinates (Appendix D.1). It enters the gradient used in the momentum update, while the decoupled weight decay keeps its coefficient λ_{WD}, so the transition matrix of the simulated system is the augmented $\widetilde{\boldsymbol{M}}_t(\eta)$ of (33).

Local simulation. Algorithm 3 first sets λ_{P} from the power-iteration estimate and then runs the local AGM dynamics (12) with the compensated gradient $\boldsymbol{g}_{\mathcal{B}_t} + (\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{P})\boldsymbol{\epsilon}_t$. The preconditioner is frozen at $\boldsymbol{P}$. The momentum starts from the checkpoint value $\boldsymbol{m}_{t_0}$ and is updated inside the simulation. The displacement starts from $\boldsymbol{\epsilon}_0 = \boldsymbol{0}$. One trajectory is run for each LR factor α_k in an increasing grid $\alpha_1 < \cdots < \alpha_K$ around the training LR η, and all trajectories use the same mini-batch sequence, fixed through the random seed s. A trajectory is marked as exploding when $\|\boldsymbol{\epsilon}_t\|_{\boldsymbol{P}}$ reaches the cutoff M at some step $t \leq T$. The estimate $\widehat{\eta}_\star$ is the smallest LR in the grid at which the trajectory explodes, and $q = \widehat{\eta}_\star/\eta$ is the critical LR ratio reported in the experiments. We

declare EoS when $q \in [1-\tau_{\text{eos}}, 1+\tau_{\text{eos}}]$, with $\tau_{\text{eos}} = 0.1$ in the experiments (Figure 4). If no factor in the grid explodes, the run is recorded as censored, $q > \alpha_K$, and the checkpoint is not counted as being at EoS. We hold the training LR η fixed during the simulation because the simulation horizon T is short relative to the training schedule. This approximates schedules such as WSD or cosine whenever their LR changes little over the simulation window.

The population detector. The analysis concerns a population version of this detector. Fix a search window $\mathcal{I}$ around η, the horizon T and a cutoff $M > 1$. Run the same compensated dynamics from every unit displacement $\|\boldsymbol{\epsilon}_0\| = 1$ with the checkpoint momentum, and take the expectation over the mini-batches. The population detector

$$\widehat{\eta}_\star(M,T) = \sup\Big\{\eta' \in \mathcal{I} : \max_{0\le t\le T} \mathbb{E}\|\boldsymbol{\epsilon}_t\|^2 \le M^2 \text{ for every } \|\boldsymbol{\epsilon}_0\| = 1\Big\}$$

is the largest LR in the window that is accepted (Appendix D.4). The implemented detector differs from it in three ways. First, the population detector takes the supremum of accepted LRs in a continuous window, while the algorithm returns the first exploding LR in an increasing grid. Finite-horizon acceptance need not be monotone in the LR, so these thresholds can differ. Second, the population detector starts every run from a unit displacement and uses a relative cutoff $M > 1$, while the algorithm starts from zero displacement under the gradient forcing and uses an absolute cutoff. Third, the population detector takes the expectation of $\|\boldsymbol{\epsilon}_t\|^2$ over the mini-batch sequence, while the algorithm runs one trajectory. The cutoff uses the norm of the frozen preconditioner because the compensation term is measured in that norm. For the population detector, the finite-horizon bound on the moment radius (Lemma G.9) is unchanged when the cutoff and the unit initialization both use this norm (Remark G.10).

Guarantee. Compensation changes the critical LR, and simulating for only T steps introduces a second error. The following theorem bounds the two contributions separately, in the inverse LR scale used by the formal result. It assumes the frozen compensated model of Assumption D.14, including $\boldsymbol{A} + \lambda_{\mathsf{P}}\boldsymbol{I} \succ 0$, and a search window with $\eta'\lambda_{\text{WD}} \le 1$ for every candidate LR η'. It also requires the population detector to accept the compensated critical LR (Assumption D.15).

Theorem B.5 (Informal; compensated finite-horizon threshold)**.** *Let $\eta_\star$ be the projected critical LR of AGM and $\widehat{\eta}_\star(M,T)$ the population detector above. Let v bound the Hessian-noise second moment by v times the mean effective curvature in both the projected and compensated systems, as in (82). Under the assumptions in Appendix D.4,*

$$\left|\frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star}\right| \le \underbrace{\frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}}}_{\textit{compensation}} + \underbrace{\frac{v}{2}}_{\textit{Hessian noise}} + O\Big(\frac{1+\log M}{T}\Big).$$

The $O(\cdot)$ is for a fixed model and search window as $(1+\log M)/T \to 0$, over choices of M and T that satisfy Assumption D.15.

The compensation term is linear in λ_{P}, which equals $\widehat{\nu}\,\widehat{\lambda}_{\max}$ for the detector. The noise level v need not be small. For a fixed cutoff M, the finite-horizon term decreases as $1/T$. Corollary G.8 proves this simplified bound. Theorem D.16 gives a sharper comparison that retains the shared noise. Small saddleness alone need not make the total error small. The theorem concerns the expectation over mini-batches from unit initial displacements, not the single trajectory from zero displacement that the experiments run. The estimated shift λ_{P} alone does not guarantee $\boldsymbol{A} + \lambda_{\mathsf{P}}\boldsymbol{I} \succ 0$. Proposition G.11 gives a sufficient cutoff for accepting the compensated critical LR. The complete assumptions and sharper theorem are in Appendix D.4. The simplified bound and its proof are in Appendix G.3.

B.4 Cost of a Hessian-Vector Product

Curvature enters both procedures only through Hessian-vector products (HVPs), which we compute with PyTorch's autograd by differentiating a gradient-vector inner product,

$$\boldsymbol{g}(\boldsymbol{w}) = \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}), \qquad \text{HVP}(\boldsymbol{v}) = \nabla_{\boldsymbol{w}}\big(\boldsymbol{g}(\boldsymbol{w})^\top \boldsymbol{v}\big) = \nabla^2_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w})\,\boldsymbol{v}. \tag{21}$$

We estimate the cost for a linear layer, since matrix multiplications dominate the attention and MLP projections in our models.

Matrix-multiplication count for a linear layer. Consider a linear layer

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}, \tag{22}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times d_{\text{in}}}$ holds a batch of n inputs, $\boldsymbol{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, and $\boldsymbol{Y} \in \mathbb{R}^{n \times d_{\text{out}}}$. Let $\boldsymbol{G} = \partial \mathcal{L} / \partial \boldsymbol{Y}$ be the upstream gradient. The forward pass uses one general matrix multiplication (GEMM),

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W} \quad \Rightarrow \quad 1 \text{ GEMM}. \tag{23}$$

The first-order backward pass computes gradients with respect to $\boldsymbol{X}$ and $\boldsymbol{W}$,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{X}} = \boldsymbol{G}\boldsymbol{W}^\top, \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = \boldsymbol{X}^\top \boldsymbol{G}, \tag{24}$$

which costs two GEMMs. Thus, a forward and backward pass through a linear layer costs approximately three GEMMs, ignoring elementwise operations.

The autograd procedure above computes an HVP by backpropagating through the gradient computation in a second reverse-mode pass. For layers whose cost is dominated by matrix multiplication, this incurs overhead comparable to an additional forward and backward pass through the same linear operators, plus extra GEMMs from differentiating the backward equations. Including the required second-order terms gives approximately

$$\underbrace{1}_{\text{forward}} + \underbrace{2}_{\text{backward}} + \underbrace{2}_{\text{differentiate backward (w.r.t. } \boldsymbol{W})} + \underbrace{4}_{\text{propagate directional adjoints}} \approx 9 \text{ GEMMs}. \tag{25}$$

This is about $3\times$ the cost of one gradient evaluation (9 versus 3 GEMMs per linear layer), or $9\times$ a forward pass when GEMMs dominate. Since transformer blocks contain many such linear layers (QKV projections, the attention output projection, and the two MLP projections), this constant-factor overhead provides a practical estimate of HVP cost for our models.

C Experimental Details

This appendix gives the models, training and detection settings behind Section 5, the protocol of the batch size threshold experiment, and the figures referenced from the main text but not shown there.

C.1 Models, Training, and Detection Settings

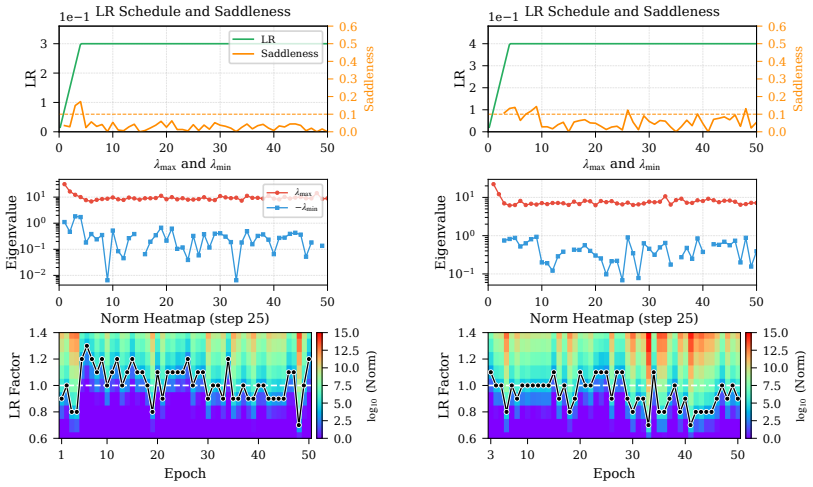
Model architectures. Table 1 lists the models used in our experiments.

Table 1: Model architectures used in our experiments.

Model	#Layers	Hidden Dim	#Heads	Head Dim	Params
GPT-extra-small	4	768	12	64	43M
GPT-small	12	768	12	64	104M
GPT-medium	24	1,024	16	64	344M
GPT-large	36	1,280	20	64	787M

For each dataset, we train a byte-level BPE tokenizer with a vocabulary of $8{,}192$ tokens and use it for all model sizes. We use a small vocabulary instead of the standard $50{,}257$-token GPT-2 tokenizer because rarely trained tokens distort the preconditioned curvature. Under Adam, the second-moment estimate of an embedding row grows with how often its token is trained on, so a token that is seldom seen has $\sqrt{\hat{v}}$ close to the floor set by ϵ. When such a token appears in the batch used for the Hessian-vector products, its row is amplified by roughly $1/\sqrt{\hat{v}}$ and can dominate the top eigenvalues. A smaller vocabulary leaves fewer such tokens. With $8{,}192$ tokens, at most 72 tokens are rare by the criterion below at the checkpoints we used to build the filter.

Even with the smaller vocabulary, we remove the remaining rare tokens from the data used for EoS and EoC detection. All detection batches come from a single training shard that no run is trained on. For a run with batch size B at checkpoint step K, we call a token rare if it has not appeared in the first K training steps, if it has appeared less than once per step on average, or if it has not appeared



(a) ResNet20 Model (b) ResNet32 Model

Figure 7: **EoS and EoC appears on model trained with SGD.** We train ResNet20 and ResNet32 with MSE loss and SGD, using batch size 1024 and weight decay 5×10^{-4} for 50 epochs. Panels follow Figure 2, using the raw Hessian and epoch on the horizontal axis.

in the last 20 steps. The padding token is always treated as rare. We take the union of the rare sets over the stable-phase checkpoints of all three model sizes (72 tokens) and drop every sequence of the shard whose inputs or targets contain a rare token. We also drop every sequence in which a token that appears fewer than 100 times per training step occurs more than 32 times, because such a burst over-represents that token's row in a single batch. The filtered shard keeps 92,060 of 97,656 sequences (94.3%), and the same shard is used for every model size, learning rate and checkpoint.

For model architectures, we apply QK-normalization (Henry et al., 2020) and attention gating (Qiu et al., 2025) for stability. All models are trained with a sequence length of 1024.

We run most experiments on GPT-extra-small. Unless otherwise specified, we use 100 stochastic power iterations and 50 compensated local simulation steps for GPT-extra-small. For the larger models (GPT-small/medium/large), we use 50 power iterations and 25 local simulation steps. In all local simulations, we mark a run as exploding when the displacement norm reaches the cutoff $M = 10^5$.

C.2 Additional Results

EoS across the pretraining runs. The bounded-to-exploding transition of the compensated local simulation appears throughout our AdamW pretraining runs. Across models from 43M to 787M parameters, Figure 2 shows many checkpoints at which the local simulation changes sharply from bounded trajectories to explosion near LR factor 1. Figures 3 and 8 show that the location and duration of this behavior depend on the LR schedule and the peak LR. Figures 4 and 9 show that it also depends on the batch size. Taken together, these experiments show that EoS is common in our AdamW pretraining runs, and absent at the smallest batch size and extreme peak LRs. Within the LR range tested around the training LR, the simulations show one transition at most checkpoints. Under the assumptions of Theorem D.7, strict homogeneous stability holds on a single LR interval whose endpoint is the AGM critical LR (Proposition B.3).

Robustness to the compensation coefficient. Figure 10 shows one checkpoint with three compensation coefficients. The sharp transition is absent without compensation, appears at the estimated value, and remains visible at the larger value 600.

C.3 Batch Size Threshold Experiments

This subsection gives the protocol behind Section 5.3. We study five datasets (FineWeb (Penedo et al., 2024), OpenWebMath (Paster et al., 2024), OpenWebText (Gokaslan & Cohen, 2019), C4 (Raffel et al., 2020), and DCLM (Li et al., 2024)) using GPT-extra-small. We use AdamW and the WSD schedule throughout, with baseline weight decay $\lambda_{\mathrm{WD}} = 0.1$.

To vary the batch size, we follow the square-root scaling rule (Malladi et al., 2024). When increasing the batch size from B to kB, we scale the LR and the weight decay, keeping the remaining hyperparameters fixed,

$$\eta \leftarrow \sqrt{k}\,\eta, \qquad \lambda_{\mathrm{WD}} \leftarrow \sqrt{k}\,\lambda_{\mathrm{WD}},$$

so that the timescale $B/(\eta\lambda_{\mathrm{WD}})$ is the same for all batch sizes. We also adjust the second-moment decay coefficient β_2 following Marek et al. (2025):

$$\beta_2 \leftarrow \beta_2^k.$$

CBZ estimation. For each dataset, we sweep the batch size under this scaling rule and record the final evaluation loss of each run. We estimate the CBZ discussed in Section 5.3 as the largest batch size whose final evaluation loss stays within 0.3% of the best loss over the sweep.

EBZ estimation. For each batch size B, we estimate the critical LR ratio $q = \widehat{\eta}_\star/\eta$ defined in Section 5.3, averaged over checkpoints. Runs whose local simulations stay bounded over the whole LR-factor grid are recorded as censored at the top of the grid. We then fit the averaged ratio as a function of B. The EBZ of Section 5.3 is the smallest B at which the fitted curve reaches 1.1.

Figure 4a compares EBZ to CBZ. Across the five datasets, EBZ and CBZ are strongly positively correlated (Pearson correlation 0.95). EBZ is below CBZ for C4 and FineWeb and above it for OpenWebMath, DCLM, and OpenWebText, so neither threshold is uniformly smaller.

C.4 Multistep LR Drop Ablations

Abrupt LR drops. We evaluate multistep schedule variants in Figure 11. Before the drop, all six runs are at EoS ($0.9 \le q \le 1.1$) at most checkpoints, 8 of the 9 checkpoints before the drop in panels a–c and 12 to 13 of the 14 in panels d–f. After the drop, the transition moves to larger LR factors, and EoS returns only when the LR after the drop stays large. With the peak LR fixed at 0.002, no checkpoint after the drop is at EoS for any of the drop ratios 0.6, 0.45, and 0.3 (panels a–c). The median q after the drop is 1.3 for the ratio 0.6 and 1.4 for 0.45. For the ratio 0.3, no LR factor in the grid explodes at 35 of the 41 checkpoints after the drop. With the drop ratio fixed at 0.7 (panels d–f), EoS returns at 26 of the 36 checkpoints after the drop for the peak LR 0.008, at 7 for 0.004, and at none for 0.0005.

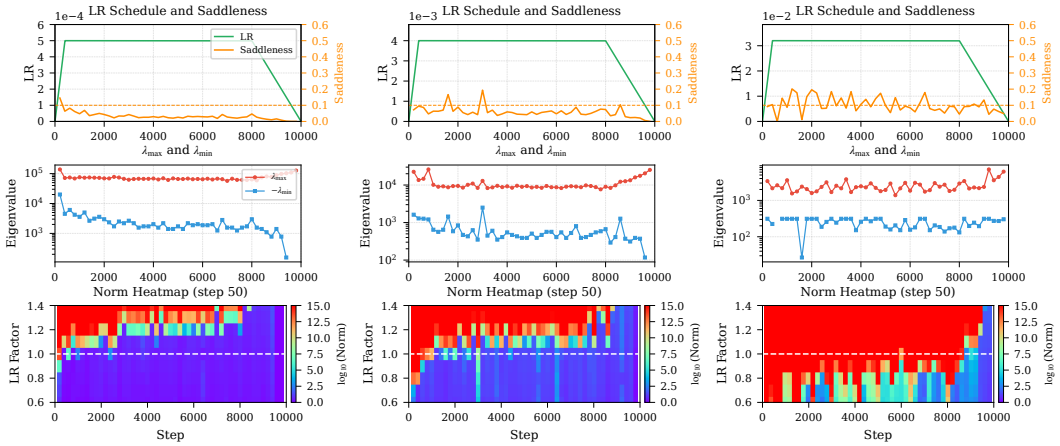
C.5 Linear LR Decay Ablations

We evaluate different linear-decay configurations in Figure 12. As in the multistep ablations, the decay ratio affects whether and when EoS re-emerges during the decay phase. The transition time also depends on the decay length. Longer decay schedules give a slower transition, with the transition time roughly proportional to the number of decay steps in our settings.

Figure 12 shows how EoC changes during LR decay. As the LR decays, $-\widehat{\lambda}_{\min}$ falls from about 850 in the stable phase to between 56 and 540 once the LR is below half its peak, and the saddleness falls from about 0.09 to 0.03 or less. EoC therefore weakens in the small-LR regime of these runs.

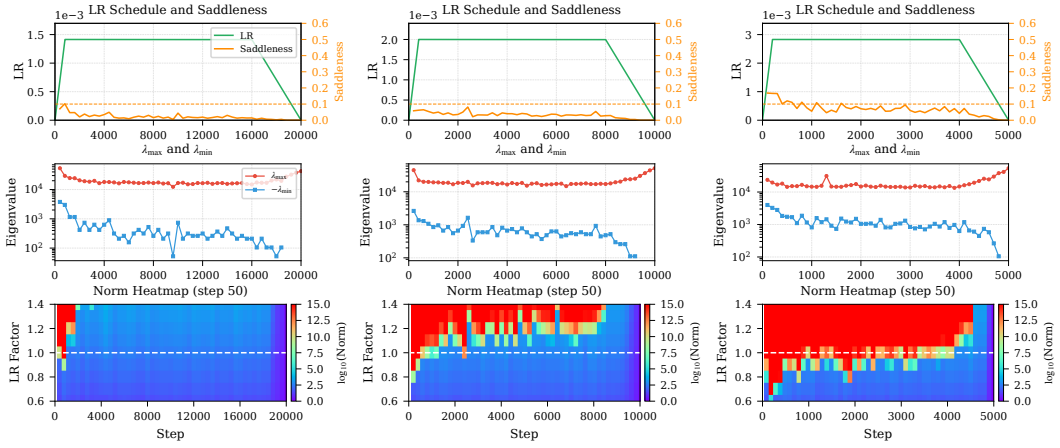
C.6 Experiments Across Model Scales

We train GPT-small, GPT-medium, and GPT-large (Table 1) on FineWeb for 20 tokens per parameter with batch sizes 512, 1,024, and 1,536, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1), and a WSD schedule with 4% warmup and a linear decay to zero over the last 20% of steps. For each size, we sweep eight peak LRs from 3.125×10^{-4} to 0.04, doubling at each step. We run EoS and EoC detection at ten checkpoints in the stable phase of each run, with compensated local simulations at LR factors from 0.6 to 1.4. For the runs with large peak LRs, where no LR factor from 0.6 stays bounded



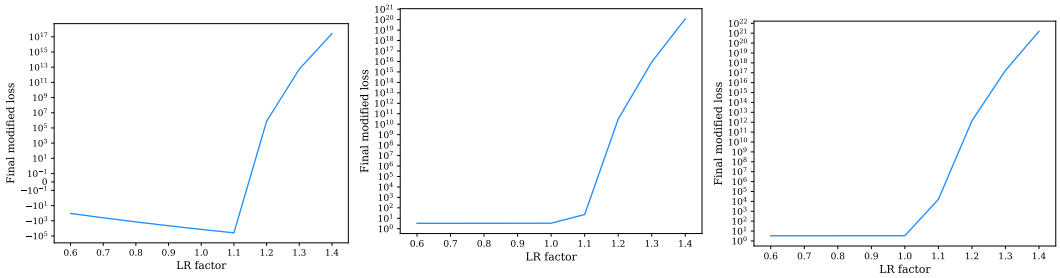
(a) WSD, LR $= 0.0005$. (b) WSD, LR $= 0.004$. (c) WSD, LR $= 0.032$.

Figure 8: **EoS and EoC measurements under different LRs.** We train GPT-extra-small (43M) on FineWeb-10B using AdamW with the WSD schedule and vary the LR. All runs use 400 linear warmup steps, 7600 stable steps, and 2000 linear decay steps. We identify 0.004 as the optimal peak LR via a separate sweep.



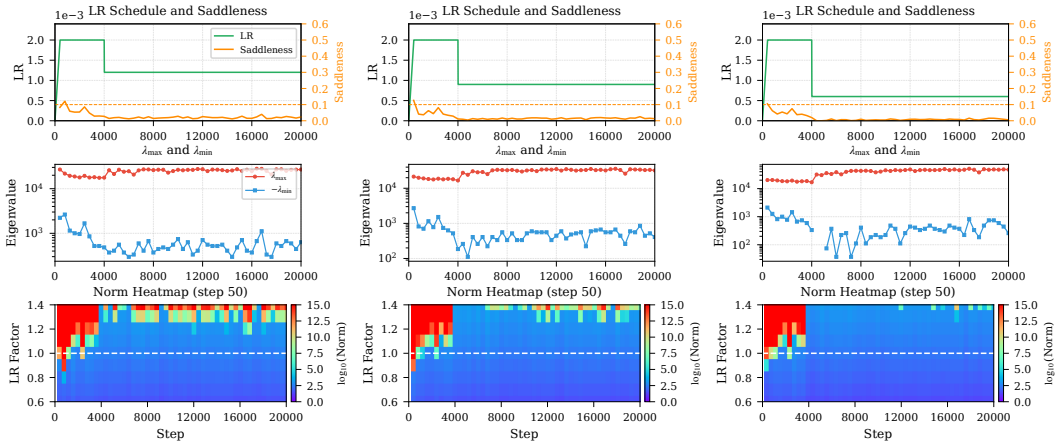
(a) WSD, batch size $= 256$. (b) WSD, batch size $= 512$. (c) WSD, batch size $= 1024$.

Figure 9: **EoS emerges only above a batch size threshold.** We train GPT-extra-small (43M) on FineWeb using AdamW with the WSD schedule at global batch sizes 256, 512, and 1,024 with the same number of training tokens. Starting from LR 0.002, weight decay 0.1, and $\beta_2 = 0.95$ at batch size 512, we scale the LR, the weight decay, and β_2 with the square-root rule of Appendix C.3. Each panel follows the reporting protocol of Figure 2. (a) At batch size 256, the transition is near factor 1.0 only at the first checkpoints, and from step 3,600 on no LR factor in the grid explodes. (b, c) At 512 and 1,024, the transition stays close to factor 1.0 for most of training.

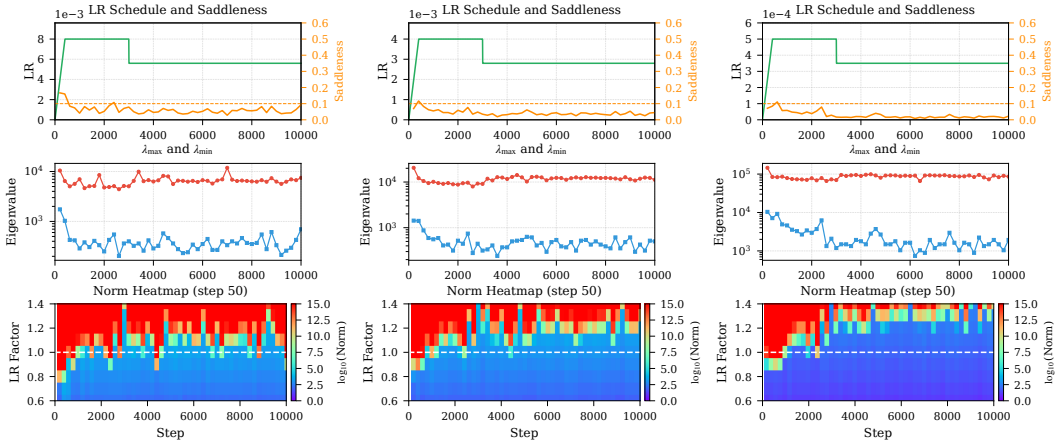


(a) No compensation ($\lambda_P = 0$). (b) Estimated compensation ($\lambda_P \approx 408$). (c) Compensation $\lambda_P = 600$.

Figure 10: **Effects of different compensation coefficients.** We train GPT-extra-small (43M) with AdamW on FineWeb-10B, run EoS detection with different compensation coefficients λ_P, and report the loss at local step 25. (a) $\lambda_P = 0$ (no compensation). (b) $\lambda_P \approx 408$ (estimated). (c) $\lambda_P = 600$.

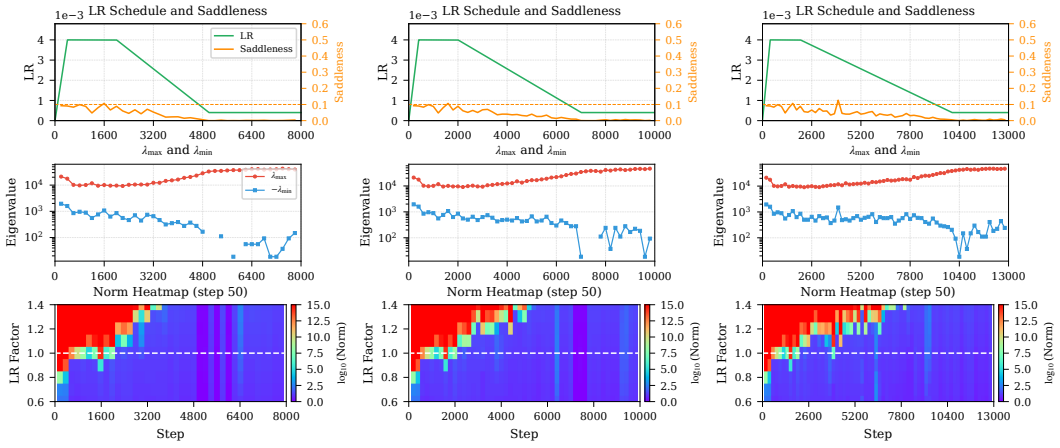


(a) Multistep, drop ratio = 0.6. (b) Multistep, drop ratio = 0.45. (c) Multistep, drop ratio = 0.3.

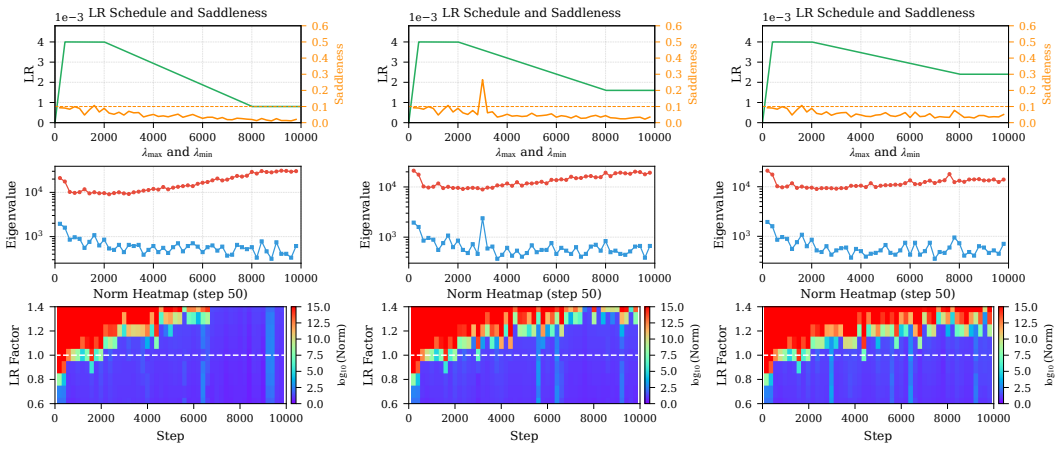


(d) Multistep, peak LR = 0.008. (e) Multistep, peak LR = 0.004. (f) Multistep, peak LR = 0.0005.

Figure 11: **Ablations on the multistep LR drop.** We train GPT-extra-small (43M) and vary the peak LR and the drop ratio of the multistep schedule. (a)–(c) The peak LR is fixed at 0.002 and the drop ratio varies over 0.6, 0.45, and 0.3, with the drop at step 4000. (d)–(f) The drop ratio is fixed at 0.7 and the peak LR varies over 0.008, 0.004, and 0.0005, with the drop at step 3000.



(a) Decay length = 3000. (b) Decay length = 5000. (c) Decay length = 8000.



(d) Decay ratio = 0.2. (e) Decay ratio = 0.4. (f) Decay ratio = 0.6.

Figure 12: **Ablations for the linear decay phase.** We train GPT-extra-small (43M) and vary the decay length and decay ratio of the linear decay phase. (a)–(c) The decay ratio is fixed at 0.1, with decay lengths of 3000, 5000, and 8000 steps. (d)–(f) The decay length is fixed at 6000 steps, with decay ratios of 0.2, 0.4, and 0.6.

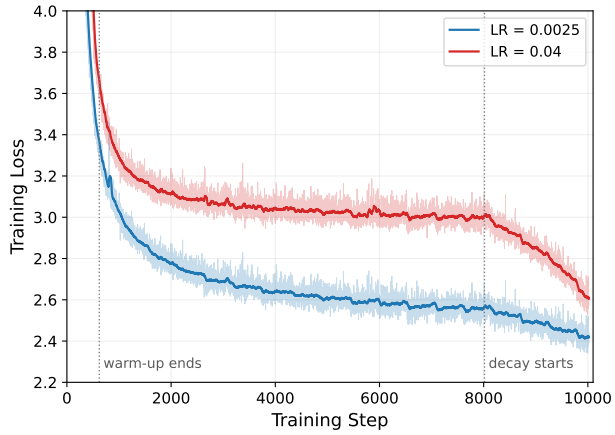


Figure 13: **Training loss of GPT-large.** We compare training loss versus steps at the empirically optimal peak LR (0.0025) and the largest tested peak LR (0.04). Both runs are trained for 20 tokens per parameter.

at many checkpoints, we re-estimate λ_P at LR factor 0.1 and add LR factors 0.2 to 0.5. Figures 14 to 16 show all 24 runs, and Figure 5 summarizes them. Figure 2 uses one additional run per size at peak LR 1.25×10^{-3}, with checkpoints saved throughout training.

At the three smallest peak LRs, 17, 18, and 19 of the 30 checkpoints of each size are at EoS ($0.9 \leq q \leq 1.1$), and 26, 27, and 26 satisfy the EoC criterion. From peak LR 0.01 on, only 12 of the 90 checkpoints are at EoS, and at 38 of them no LR factor in the grid stays bounded.

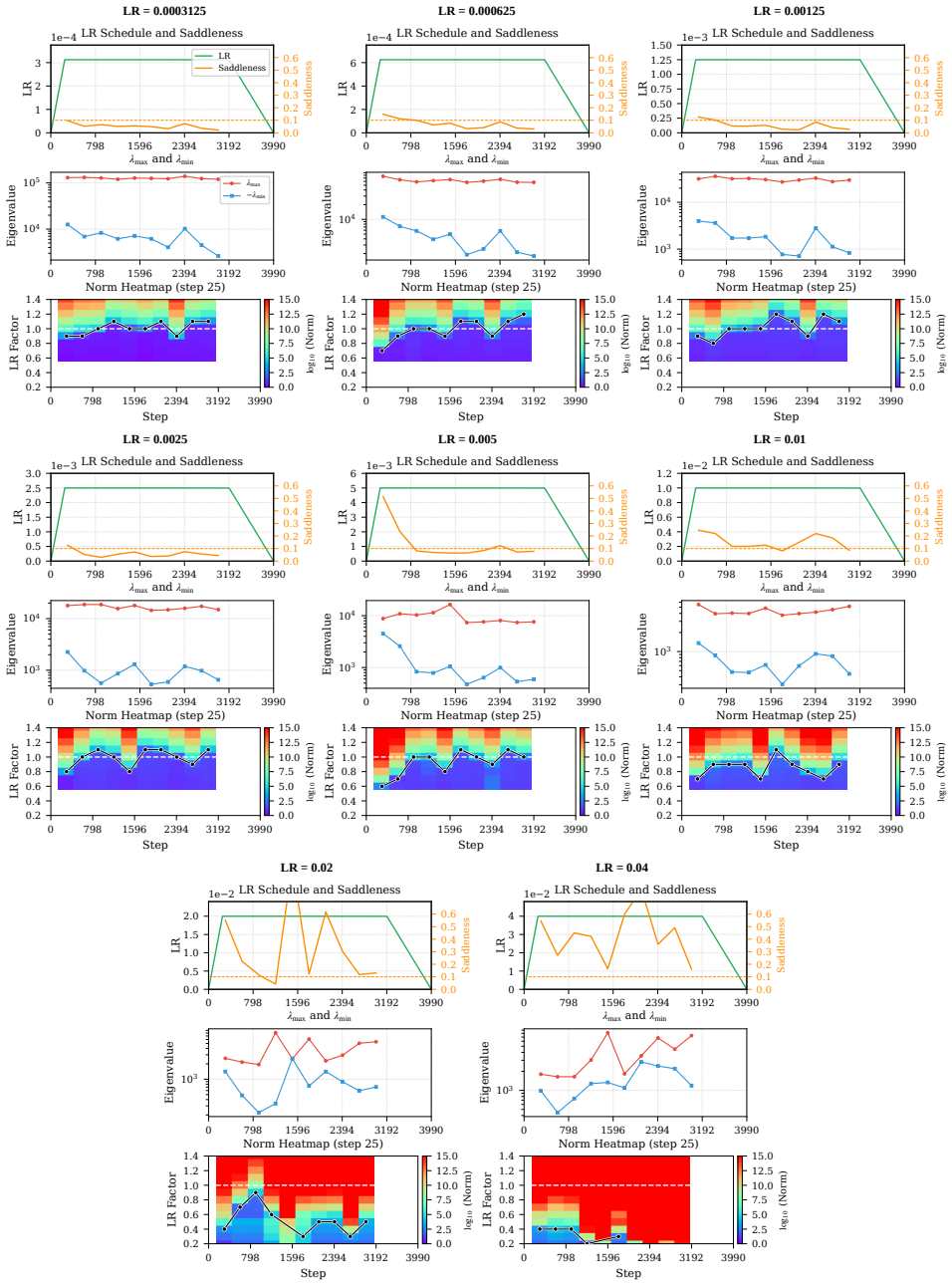


Figure 14: **GPT-small training stays near EoS up to peak LR** 0.01 **and goes beyond EoS at larger peak LRs.** Up to peak LR 0.01, the critical LR factor $q = \widehat{\eta}_\star/\eta$, the smallest LR factor whose simulation explodes, lies between 0.9 and 1.2 at 51 of the 60 checkpoints, and 35 of these are at EoS ($0.9 \leq q \leq 1.1$). Over the same peak LRs, 44 of the 60 checkpoints satisfy the EoC criterion. For peak LRs from 0.02 on, the training LR exceeds the estimated critical LR ($q < 0.9$) at 19 of the 20 checkpoints, and at 5 of the 10 checkpoints of peak LR 0.04 no LR factor in the grid stays bounded. In these two runs, the rows below LR factor 0.6, the $-\lambda_{\min}$ curve, and the saddleness use the compensation coefficient λ_{P} estimated at LR factor 0.1, while the rows from 0.6 up use the one estimated at 0.6. Because the two coefficients differ, a bounded factor 0.6 sits above exploding lower factors at one checkpoint of peak LR 0.02. Each panel is one peak LR, each heatmap column is one of ten checkpoints in the stable phase, and the black line marks the largest LR factor that stays bounded. Saddleness above 0.7 runs off the top of the axis, and blank heatmap cells were not simulated. See Figure 2 for the panel format and Appendix C.6 for the setup.

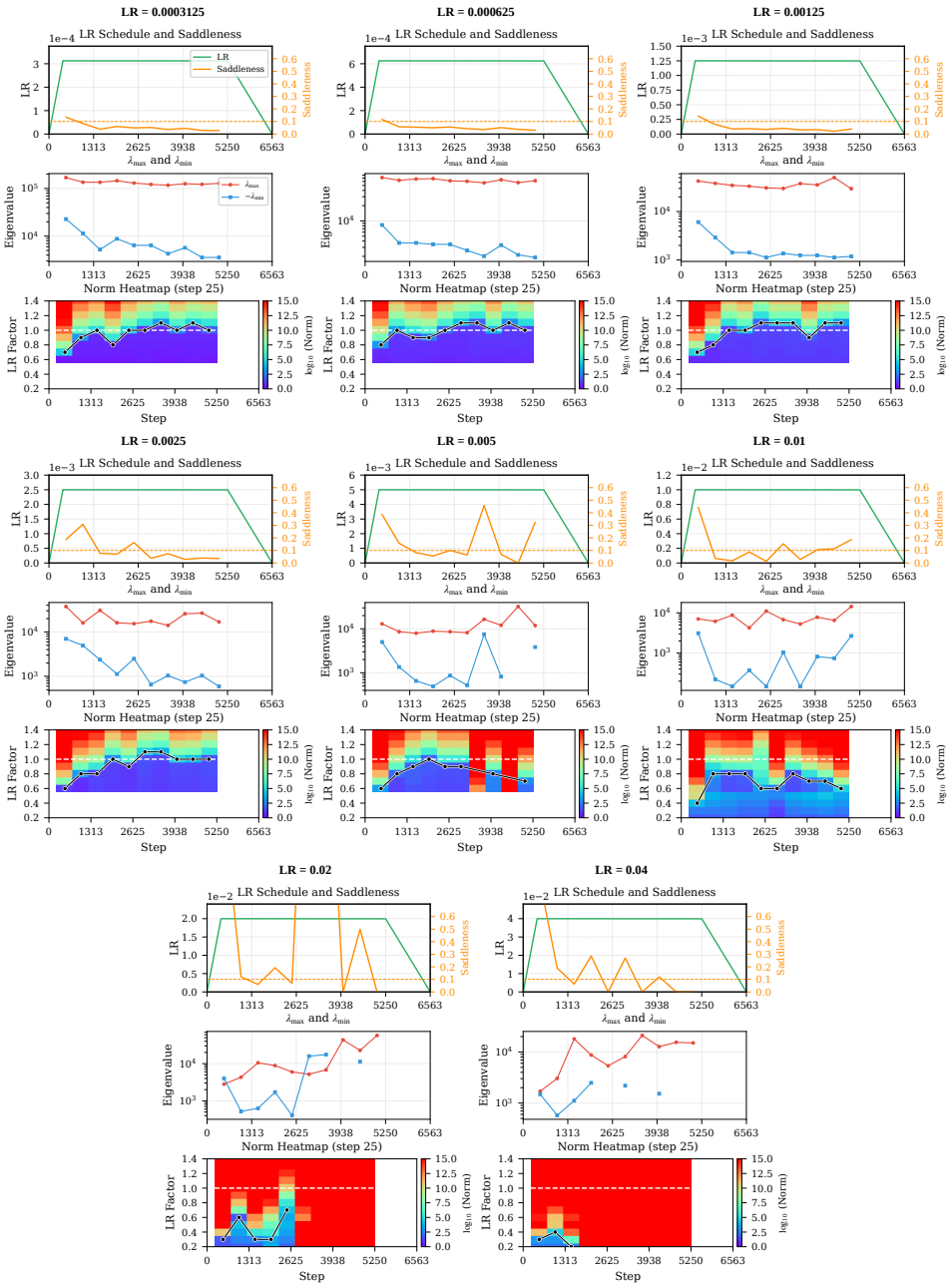


Figure 15: **GPT-medium training stays near EoS up to peak LR 5×10^{-3} and goes beyond EoS at larger peak LRs.** Up to peak LR 5×10^{-3}, the critical LR factor $q = \widehat{\eta}_\star/\eta$, the smallest LR factor whose simulation explodes, lies between 0.9 and 1.2 at 43 of the 50 checkpoints, and 31 of these are at EoS ($0.9 \leq q \leq 1.1$). Over the same peak LRs, 38 of the 50 checkpoints satisfy the EoC criterion. For peak LRs from 0.01 on, the training LR exceeds the estimated critical LR ($q < 0.9$) at 26 of the 30 checkpoints, and at 7 of the 10 checkpoints of peak LR 0.04 no LR factor in the grid stays bounded. In these three runs, the rows below LR factor 0.6, the $-\lambda_{\min}$ curve, and the saddleness use the compensation coefficient λ_{P} estimated at LR factor 0.1, while the rows from 0.6 up use the one estimated at 0.6. Each panel is one peak LR, each heatmap column is one of ten checkpoints in the stable phase, and the black line marks the largest LR factor that stays bounded. Saddleness above 0.7 runs off the top of the axis, and blank heatmap cells were not simulated. See Figure 2 for the panel format and Appendix C.6 for the setup.

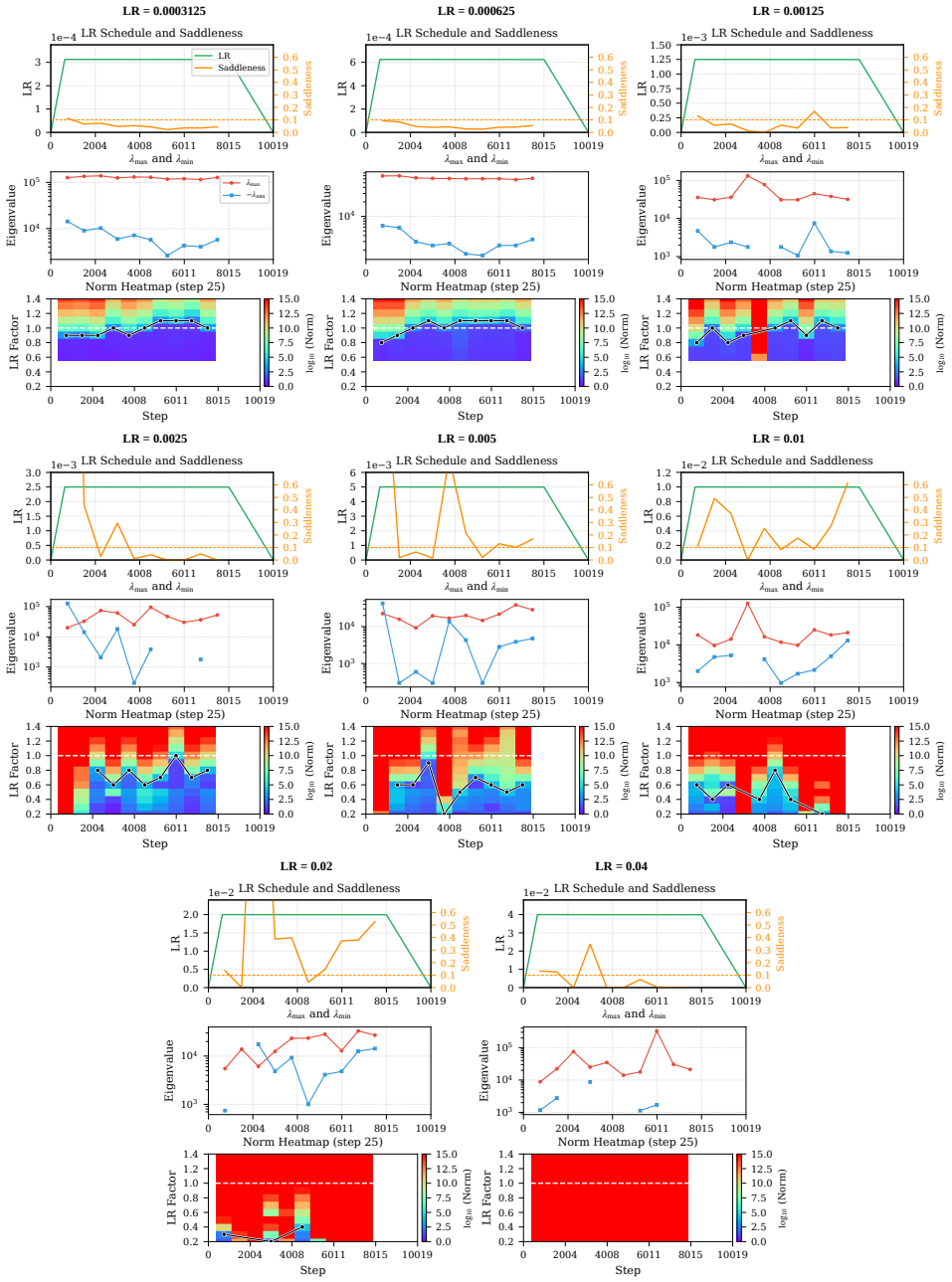


Figure 16: **GPT-large training stays near EoS up to peak LR 1.25×10^{-3} and goes beyond EoS at larger peak LRs.** Up to peak LR 1.25×10^{-3}, the critical LR factor $q = \widehat{\eta}_\star/\eta$, the smallest LR factor whose simulation explodes, lies between 0.9 and 1.2 at 29 of the 30 checkpoints, and 19 of these are at EoS ($0.9 \leq q \leq 1.1$). Over the same peak LRs, 26 of the 30 checkpoints satisfy the EoC criterion. For peak LRs from 2.5×10^{-3} on, the training LR exceeds the estimated critical LR ($q < 0.9$) at 44 of the 50 checkpoints, and at every checkpoint of peak LR 0.04 no LR factor in the grid stays bounded. In these five runs, the rows below LR factor 0.6, the $-\lambda_{\min}$ curve, and the saddleness use the compensation coefficient λ_P estimated at LR factor 0.1, while the rows from 0.6 up use the one estimated at 0.6. Because the two coefficients differ, a bounded factor 0.6 sits above exploding lower factors at two checkpoints of peak LR 2.5×10^{-3} and one of 0.01. Each panel is one peak LR, each heatmap column is one of ten checkpoints in the stable phase, and the black line marks the largest LR factor that stays bounded. Saddleness above 0.7 runs off the top of the axis, and blank heatmap cells were not simulated. See Figure 2 for the panel format and Appendix C.6 for the setup.

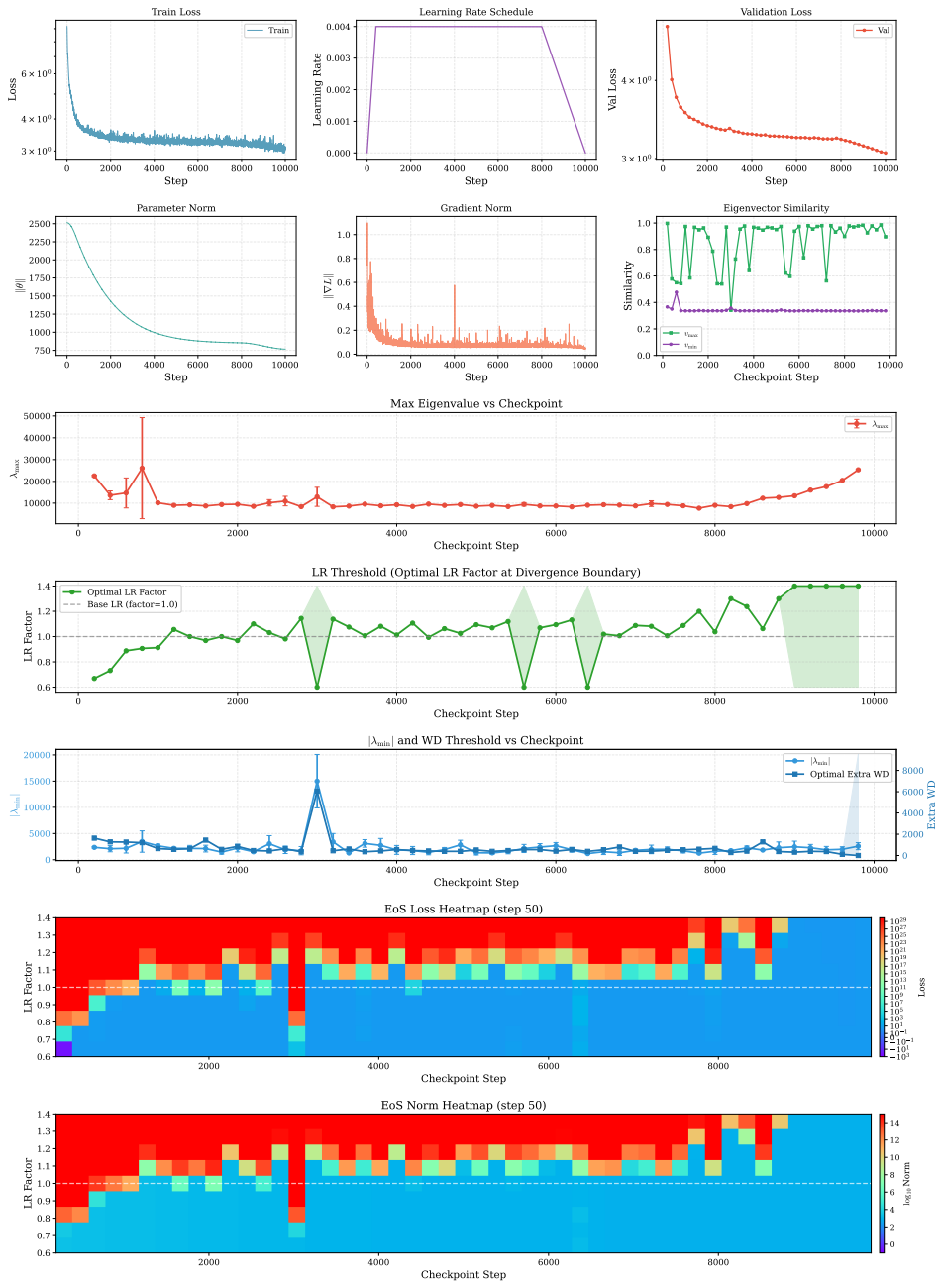


Figure 17: **Summary of a representative training run.** We show training dynamics for GPT-extra-small trained on FineWeb-10B using AdamW with LR 0.004, weight decay 0.1, and the WSD LR schedule.

D THEORETICAL ANALYSIS

This appendix gives the local moment recursions, stability thresholds, and detector error bounds. Proofs are in Appendices E to G.

D.1 SETTING AND THE LOCAL LINEAR SYSTEM

The local system. We consider adaptive gradient methods (AGMs) (Li et al., 2025), including Adam (Kingma, 2014) and AdamW (Loshchilov & Hutter, 2017), with the updates of (AGM), recalled here,

$$\boldsymbol{m}_{t+1} = \beta_1 \cdot \boldsymbol{m}_t + (1-\beta_1) \cdot \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_t) \tag{26}$$

$$\boldsymbol{P}_{t+1} = \Psi(\boldsymbol{P}_t, \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_t), \boldsymbol{w}_t) \tag{27}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{P}_{t+1}^{-1} \boldsymbol{m}_{t+1} - \eta \lambda_{\mathrm{WD}} \boldsymbol{w}_t, \tag{28}$$

where $\boldsymbol{m}_t \in \mathbb{R}^d$ is the momentum and Ψ maps a positive definite preconditioner and two vectors in $\mathbb{R}^d$ to a positive definite preconditioner. This class includes SGD with and without momentum. We assume that the preconditioner varies slowly along the training trajectory and freeze it over the short window of a local simulation at a checkpoint $\boldsymbol{w}_{t_0}$, writing $\boldsymbol{P} \coloneqq \boldsymbol{P}_{t_0}$. This is an idealization. Adam's second-moment decay coefficient β_2 close to one motivates this approximation, but freezing $\boldsymbol{P}_t$ still requires its change over the simulation window to be small.

Recall the mini-batch quadratic surrogate (11) from Appendix B.1:

$$\begin{aligned} Q_{\mathcal{B}_t}(\boldsymbol{w}) \coloneqq \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) + (\boldsymbol{w}-\boldsymbol{w}_{t_0})^\top \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) \\ + \tfrac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_{t_0})^\top \nabla^2 \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0})(\boldsymbol{w}-\boldsymbol{w}_{t_0}), \end{aligned} \tag{29}$$

where $\mathcal{B}_t$ is the mini-batch sampled at local step t. SGD with LR η on this surrogate gives

$$\boldsymbol{w}_{t_0+t+1} = \boldsymbol{w}_{t_0+t} - \eta \nabla Q_{\mathcal{B}_t}(\boldsymbol{w}_{t_0+t}). \tag{30}$$

Write $\boldsymbol{\epsilon}_t \coloneqq \boldsymbol{w}_{t_0+t} - \boldsymbol{w}_{t_0}$ for the displacement after t local steps, as in the local AGM model of Appendix B.1. For the local simulation of an AGM, the momentum and parameter updates (26) and (28) are run on the surrogate (29), with the gradient $\nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_t)$ replaced by $\nabla Q_{\mathcal{B}_t}(\boldsymbol{w}_{t_0+t}) = \boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon}_t$ and the preconditioner fixed at $\boldsymbol{P}$. Here $\boldsymbol{g}_{\mathcal{B}_t} = \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0})$ and $\boldsymbol{H}_{\mathcal{B}_t} = \nabla^2 \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0})$ are the mini-batch gradient and Hessian at the checkpoint, as in Section 4. The local dynamics are

$$\begin{aligned} \boldsymbol{m}_{t+1} &= \beta_1 \boldsymbol{m}_t + (1-\beta_1)(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon}_t), \\ \boldsymbol{\epsilon}_{t+1} &= \boldsymbol{\epsilon}_t - \eta \boldsymbol{P}^{-1}\boldsymbol{m}_{t+1} - \eta \lambda_{\mathrm{WD}}(\boldsymbol{\epsilon}_t + \boldsymbol{w}_{t_0}). \end{aligned} \tag{31}$$

Compensation. Compensation adds a quadratic penalty to the local surrogate, measured in the norm of the frozen preconditioner. As in Appendix B.3, we write G_t for this loss:

$$G_t(\boldsymbol{\epsilon}) \coloneqq \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) + \boldsymbol{\epsilon}^\top \nabla \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) + \tfrac{1}{2}\boldsymbol{\epsilon}^\top \nabla^2 \mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0})\boldsymbol{\epsilon} + \tfrac{\lambda_{\mathrm{P}}}{2}\|\boldsymbol{\epsilon}\|_{\boldsymbol{P}}^2.$$

Here $\|\boldsymbol{\epsilon}\|_{\boldsymbol{P}}^2 = \boldsymbol{\epsilon}^\top \boldsymbol{P}\boldsymbol{\epsilon}$, with $\boldsymbol{P}$ fixed throughout the local simulation. For SGD, $\boldsymbol{P} = \boldsymbol{I}$, so the penalty is $\frac{\lambda_{\mathrm{P}}}{2}\|\boldsymbol{\epsilon}\|^2$. Its gradient is $\nabla G_t(\boldsymbol{\epsilon}) = \boldsymbol{g}_{\mathcal{B}_t} + (\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathrm{P}}\boldsymbol{P})\boldsymbol{\epsilon}$, so the penalty replaces the mini-batch Hessian $\boldsymbol{H}_{\mathcal{B}_t}$ by $\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathrm{P}}\boldsymbol{P}$ in (31), which in preconditioned coordinates is a shift of the curvature by $\lambda_{\mathrm{P}}\boldsymbol{I}$ (Proposition D.2). This is the same shift that the Euclidean penalty in Section 4.2 applies to the raw Hessian. The decoupled weight decay keeps its coefficient λ_{WD}, and $\lambda_{\mathrm{P}} = 0$ recovers the uncompensated surrogate.

Assumption D.1 (Mini-batch model)**.** The pairs $(\boldsymbol{H}_{\mathcal{B}_t}, \boldsymbol{g}_{\mathcal{B}_t})$, $t \geq 0$, are i.i.d., independent of the initial state $(\boldsymbol{m}_0, \boldsymbol{\epsilon}_0)$, each $\boldsymbol{H}_{\mathcal{B}_t}$ is symmetric, and both have finite second moments. Their means are the population derivatives at the checkpoint,

$$\mathbb{E}[\boldsymbol{g}_{\mathcal{B}_t}] = \nabla \mathcal{L}(\boldsymbol{w}_{t_0}), \qquad \mathbb{E}[\boldsymbol{H}_{\mathcal{B}_t}] = \nabla^2 \mathcal{L}(\boldsymbol{w}_{t_0}).$$

The assumption treats the mini-batches of consecutive local steps as independent draws from one distribution. Independence across steps closes the moment recursions, and finite second moments make them well defined. Uniformly sampled batches satisfy the two mean identities, which identify the population curvature and remove the mean forcing when we center SGD at a stationary point. Nothing is assumed about the sign of the curvature. The results that need positive curvature on a subspace say so in their statements.

The augmented linear system. Centering the displacement at a reference point and changing to preconditioned coordinates turns (31) into the stochastic linear system (13) of Appendix B.1. The following proposition gives its matrices in terms of the preconditioned mini-batch Hessian $\boldsymbol{D}_{\mathcal{B}_t} := \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2}$ defined there.

Proposition D.2 (Augmented linear system)**.** *Fix a reference displacement $\boldsymbol{x}_\star \in \mathbb{R}^d$ and let*

$$\boldsymbol{z}_t = \begin{bmatrix} \eta \boldsymbol{P}^{-1/2}\boldsymbol{m}_t \\ \boldsymbol{P}^{1/2}(\boldsymbol{\epsilon}_t - \boldsymbol{x}_\star) \end{bmatrix}, \qquad \boldsymbol{h}_t = (1-\beta_1)\eta \boldsymbol{P}^{-1/2}(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{x}_\star).$$

Under the local dynamics (31), $\boldsymbol{z}_{t+1} = \boldsymbol{M}_t(\eta)\boldsymbol{z}_t + \boldsymbol{q}_t$, where

$$\boldsymbol{M}_t(\eta) = \begin{bmatrix} \beta_1 \boldsymbol{I} & (1-\beta_1)\eta \boldsymbol{D}_{\mathcal{B}_t} \\ -\beta_1 \boldsymbol{I} & (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{I} - (1-\beta_1)\eta \boldsymbol{D}_{\mathcal{B}_t} \end{bmatrix}, \quad \boldsymbol{q}_t = \begin{bmatrix} \boldsymbol{h}_t \\ -\boldsymbol{h}_t - \eta\lambda_{\mathrm{WD}}\boldsymbol{P}^{1/2}(\boldsymbol{x}_\star + \boldsymbol{w}_{t_0}) \end{bmatrix}. \tag{32}$$

With compensation, the same holds with $\boldsymbol{D}_{\mathcal{B}_t}$ replaced by $\boldsymbol{D}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I}$ in $\boldsymbol{M}_t(\eta)$ and $\boldsymbol{H}_{\mathcal{B}_t}$ replaced by $\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{P}$ in $\boldsymbol{h}_t$. The transition matrix becomes

$$\widetilde{\boldsymbol{M}}_t(\eta) = \begin{bmatrix} \beta_1 \boldsymbol{I} & (1-\beta_1)\eta(\boldsymbol{D}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I}) \\ -\beta_1 \boldsymbol{I} & (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{I} - (1-\beta_1)\eta(\boldsymbol{D}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I}) \end{bmatrix}. \tag{33}$$

Under Assumption D.1, $(\boldsymbol{M}_t(\eta), \boldsymbol{q}_t)$ is a function of $\mathcal{B}_t$ alone and is independent of $\boldsymbol{z}_t$.

The reference point enters only the forcing $\boldsymbol{q}_t$. The homogeneous system $\boldsymbol{z}_{t+1} = \boldsymbol{M}_t(\eta)\boldsymbol{z}_t$, which governs the difference between two trajectories driven by the same mini-batches, does not depend on it. Compensation shifts $\boldsymbol{D}_{\mathcal{B}_t}$ by $\lambda_{\mathsf{P}}\boldsymbol{I}$ inside the momentum update only. The decoupled weight decay term is unchanged. Equation (13) in Appendix B.1 uses displacement from the checkpoint, corresponding to $\boldsymbol{x}_\star = \boldsymbol{0}$ in this proposition. The proof is in Appendix E.1.

Moment recursions and mean-square stability. When the transition and the forcing of a stochastic linear system are independent of its state, the mean and the second moment of the state obey closed recursions. We state this once, for a generic system, and apply it to the augmented system and to its SGD block.

Proposition D.3 (Moment recursion of a stochastic linear system)**.** *Let $\boldsymbol{z}_{t+1} = \boldsymbol{M}_t\boldsymbol{z}_t + \boldsymbol{q}_t$, where $\boldsymbol{z}_0$ has finite second moment and the pairs $(\boldsymbol{M}_t, \boldsymbol{q}_t)$, $t \geq 0$, are i.i.d., have finite second moments, and are independent of $\boldsymbol{z}_0$. Let $\boldsymbol{M} := \mathbb{E}[\boldsymbol{M}_t]$, $\boldsymbol{z}_t^{(1)} := \mathbb{E}[\boldsymbol{z}_t]$ and $\boldsymbol{z}_t^{(2)} := \mathsf{vec}(\mathbb{E}[\boldsymbol{z}_t\boldsymbol{z}_t^\top])$, and set*

$$\boldsymbol{S} := \mathbb{E}[\boldsymbol{q}_t \otimes \boldsymbol{M}_t + \boldsymbol{M}_t \otimes \boldsymbol{q}_t], \quad \boldsymbol{b}_{\mathrm{quad}} := \mathsf{vec}(\mathbb{E}[\boldsymbol{q}_t\boldsymbol{q}_t^\top]).$$

Then

$$\begin{bmatrix} \boldsymbol{z}_{t+1}^{(1)} \\ \boldsymbol{z}_{t+1}^{(2)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{M} & 0 \\ \boldsymbol{S} & \mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t] \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_t^{(1)} \\ \boldsymbol{z}_t^{(2)} \end{bmatrix} + \begin{bmatrix} \mathbb{E}[\boldsymbol{q}_t] \\ \boldsymbol{b}_{\mathrm{quad}} \end{bmatrix}, \tag{34}$$

and the spectral radius of the block matrix is $\max\{\rho(\boldsymbol{M}), \rho(\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t])\}$. If $\rho(\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]) < 1$, both moments converge to finite limits, so $\sup_{t\geq 0}\mathbb{E}\|\boldsymbol{z}_t\|^2 < \infty$. The transition $\boldsymbol{M}_t$ and the forcing $\boldsymbol{q}_t$ may be correlated.

Under Assumption D.1, the augmented system of Proposition D.2 is of this form with $\boldsymbol{M}_t = \boldsymbol{M}_t(\eta)$, and the compensated system with $\boldsymbol{M}_t = \widetilde{\boldsymbol{M}}_t(\eta)$. The two diagonal blocks govern the mean and the second moment. The second-moment transition splits as

$$\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t] = \boldsymbol{M} \otimes \boldsymbol{M} + \mathbb{E}[(\boldsymbol{M}_t - \boldsymbol{M}) \otimes (\boldsymbol{M}_t - \boldsymbol{M})],$$

since $\otimes$ is bilinear and $\mathbb{E}[\boldsymbol{M}_t - \boldsymbol{M}] = \boldsymbol{0}$. For the augmented system this is the AGM analogue of the decomposition

$$\mathbb{E}[\boldsymbol{M}_t^+ \otimes \boldsymbol{M}_t^+] = \mathbb{E}[\boldsymbol{M}_t^+] \otimes \mathbb{E}[\boldsymbol{M}_t^+] + \mathrm{Cov}(\boldsymbol{M}_t^+)$$

in Section 4. The mean term is the full-batch transition, and the centered term records the variation of the mini-batch Hessian. The limiting moments of the forced system need not be zero. The proof is in Appendix E.2.

We call the homogeneous system *strictly mean-square stable* if $\mathbb{E}\|\boldsymbol{z}_t\|^2 \to 0$ for every deterministic $\boldsymbol{z}_0$. By the second-moment characterization lemma (Lemma G.14) this holds if and only if $\rho(\mathbb{E}[\boldsymbol{M}_t(\eta) \otimes \boldsymbol{M}_t(\eta)]) < 1$. For the same state space, strict homogeneous stability implies bounded moments for the affine system under Proposition D.3. Comparisons of their critical LRs must also keep the retained subspace fixed, as we explain in Appendix D.2.

For SGD the displacement block of the augmented system evolves on its own, because the momentum is the current gradient. Applying Proposition D.3 to that block gives the moment recursion of Proposition 3.2 for the surrogate of a checkpoint. The stability statements follow in Appendix D.2.

Corollary D.4 (Moment recursion for SGD)**.** *Let Assumption D.1 hold, let $\beta_1 = 0$, $\boldsymbol{P} = \boldsymbol{I}$ and $\lambda_{\mathrm{WD}} = 0$, and suppose there exists a $\boldsymbol{w}_\star$ such that*

$$\nabla \mathcal{L}(\boldsymbol{w}_{t_0}) + \nabla^2 \mathcal{L}(\boldsymbol{w}_{t_0})(\boldsymbol{w}_\star - \boldsymbol{w}_{t_0}) = 0. \tag{35}$$

Under the SGD update (30), the first and second moments

$$\boldsymbol{\epsilon}_t^{(1)} := \mathbb{E}\big[\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star\big],$$
$$\boldsymbol{\epsilon}_t^{(2)} := \mathsf{vec}\Big(\mathbb{E}[(\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star)(\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star)^\top]\Big)$$

satisfy

$$\begin{bmatrix} \boldsymbol{\epsilon}_{t+1}^{(1)} \\ \boldsymbol{\epsilon}_{t+1}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[\boldsymbol{M}_t] & \boldsymbol{0} \\ \boldsymbol{S} & \mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t] \end{bmatrix} \begin{bmatrix} \boldsymbol{\epsilon}_t^{(1)} \\ \boldsymbol{\epsilon}_t^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{b}_{\mathrm{quad}} \end{bmatrix}, \tag{36}$$

where

$$\boldsymbol{M}_t := \boldsymbol{I} - \eta \boldsymbol{H}_{\mathcal{B}_t}, \qquad \boldsymbol{r}_t := \boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}(\boldsymbol{w}_\star - \boldsymbol{w}_{t_0}),$$
$$\boldsymbol{S} := -\eta\, \mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{r}_t + \boldsymbol{r}_t \otimes \boldsymbol{M}_t], \qquad \boldsymbol{b}_{\mathrm{quad}} := \eta^2 \mathsf{vec}\big(\mathbb{E}[\boldsymbol{r}_t \boldsymbol{r}_t^\top]\big),$$

and $\mathbb{E}[\boldsymbol{r}_t] = \boldsymbol{0}$.

The corollary gives closed recursions for the mean and the second moment, driven by the constant terms $\boldsymbol{b}_{\mathrm{quad}}$ and $\boldsymbol{S}$. It makes no stability claim by itself. Applying the strict spectral condition of Proposition D.3 gives the affine boundedness assertion in Proposition 3.2. The stationarity condition (35) removes the forcing from the mean recursion. It holds for some $\boldsymbol{w}_\star$ whenever $\nabla \mathcal{L}(\boldsymbol{w}_{t_0})$ lies in the range of $\nabla^2 \mathcal{L}(\boldsymbol{w}_{t_0})$, in particular whenever the Hessian is invertible. The proof is in Appendix E.3.

D.2 Stability of the Local Dynamics

Negative effective curvature makes the local dynamics unstable at every positive LR. On the strictly positive subspace, the LRs for strict homogeneous stability form a single interval.

Negative curvature. The first instability mode does not depend on the LR. Write $\boldsymbol{D} = \mathbb{E}[\boldsymbol{D}_{\mathcal{B}_t}]$ for the population preconditioned Hessian ($\boldsymbol{D}_{t_0}$ in Appendix B.1).

Lemma D.5 (Negative effective curvature)**.** *Let Assumption D.1 hold. If the effective curvature $\boldsymbol{A} = \boldsymbol{D} + \lambda_{\mathrm{WD}} \boldsymbol{I}$ has a negative eigenvalue, then the homogeneous AGM dynamics $\boldsymbol{z}_{t+1} = \boldsymbol{M}_t(\eta)\boldsymbol{z}_t$ of Proposition D.2 are not mean-square stable for any $\eta > 0$.*

Along an eigendirection of $\boldsymbol{D}$ with negative effective curvature, the mean dynamics reduce to a 2×2 system with a real eigenvalue above one for every $\eta > 0$. Momentum changes the rate of growth, not its existence. A smaller LR slows the escape and does not prevent it. For SGD the rate is explicit. Write $\boldsymbol{H}_{t_0} = \nabla^2 \mathcal{L}(\boldsymbol{w}_{t_0})$ for the population Hessian at the checkpoint.

Corollary D.6 (Negative-curvature divergence)**.** *Under the assumptions of Corollary D.4, suppose $\boldsymbol{H}_{t_0}$ has a unit eigenvector $\boldsymbol{u}$ with eigenvalue $\lambda < 0$. The population quadratic is unbounded below. For every $\eta > 0$,*

$$\boldsymbol{u}^\top \mathbb{E}[\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star] = (1 - \eta\lambda)^t\, \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{w}_{t_0} - \boldsymbol{w}_\star].$$

Thus the mean diverges whenever the initial mean has a nonzero component along $\boldsymbol{u}$. The homogeneous system is not mean-square stable.

This is the negative-curvature failure of Section 3.1, now for the surrogate of a checkpoint. The proofs of the lemma and the corollary are in Appendix E.4.

Critical LR. To isolate the instability induced by the LR, we restrict the dynamics to the subspace of positive effective curvature, as in Definition B.2. Let Π^+ be the orthogonal projection onto the strictly positive eigenspace of the symmetric effective mean $\boldsymbol{D} + \lambda_{\mathrm{WD}}\boldsymbol{I}$. All projected matrices below act on $\mathrm{range}(\Pi^+)$. Both the momentum and the displacement are restricted to this subspace, which we assume is nonempty. Directions of zero effective curvature are dropped because their mean transition has an eigenvalue equal to one, so no LR would be strictly mean-square stable if they were kept. Let $\boldsymbol{M}_t^+(\eta)$ denote the restricted transition and define

$$\boldsymbol{K}^+(\eta) \coloneqq \mathbb{E}\left[\boldsymbol{M}_t^+(\eta) \otimes \boldsymbol{M}_t^+(\eta)\right]. \tag{37}$$

The strictly mean-square stable LR set of the projected homogeneous system is

$$\mathcal{S}_{\mathrm{AGM}}(\boldsymbol{w}_{t_0}) \coloneqq \{\eta > 0 : \rho(\boldsymbol{K}^+(\eta)) < 1\}. \tag{38}$$

The retained effective mean is positive definite by construction of Π^+. The next theorem shows that under Assumption D.1 this set is the interval $(0, \eta_\star)$ for a unique finite $\eta_\star$, even when $\beta_1 > 0$. Momentum makes the transition nonsymmetric but does not create a second stable interval.

Theorem D.7 (Unique stability threshold, restated from Corollary G.4)**.** *Let Assumption D.1 hold, fix the preconditioner, $0 \le \beta_1 < 1$ and $\lambda_{\mathrm{WD}} \ge 0$, and suppose the retained effective mean is positive definite. Then the strictly mean-square stable LRs of the projected homogeneous system form a single interval $(0, \eta_\star)$ with $0 < \eta_\star < \infty$. At $\eta_\star$ the second-moment spectral radius equals one, and it is strictly greater than one for every $\eta > \eta_\star$.*

By Proposition B.3, this threshold equals the AGM critical LR defined by bounded affine moments in Definition B.2. Appendix D.4 bounds the finite-horizon error of the population detector. The same statement holds for the compensated system of Proposition D.2 whenever $\boldsymbol{D} + (\lambda_{\mathrm{WD}} + \lambda_{\mathrm{P}})\boldsymbol{I} \succ 0$, since it is the same system with a shifted effective mean. The exact stability criterion is stated in Appendix G.1. The proof of the AGM threshold statement is in Appendix G.5.3.

Above the critical LR, the second moment of some initial state grows without bound.

Corollary D.8 (Divergence above the critical LR)**.** *Under the assumptions of Theorem D.7, for every $\eta > \eta_\star$ there is a deterministic initial state $\boldsymbol{z}_0$ of the projected homogeneous system for which $\mathbb{E}\|\boldsymbol{z}_t\|^2 \to +\infty$.*

The proof is in Appendix E.5. It uses only that the second-moment spectral radius exceeds one for $\eta > \eta_\star$.

For SGD, $\beta_1 = 0$, $\boldsymbol{P} = \boldsymbol{I}$ and $\lambda_{\mathrm{WD}} = 0$, the projected system reduces to the displacement restricted to $\mathrm{range}(\Pi_{t_0}^{>0})$, where $\Pi_{t_0}^{>0}$ is the spectral projector of $\boldsymbol{H}_{t_0}$ onto its strictly positive eigenspace,

$$\boldsymbol{M}_t^+(\eta) = \boldsymbol{I} - \eta\Pi_{t_0}^{>0}\boldsymbol{H}_{\mathcal{B}_t}\Pi_{t_0}^{>0}, \qquad \boldsymbol{x}_{t+1} = \boldsymbol{M}_t^+(\eta)\boldsymbol{x}_t,$$
$$\eta_\star = \sup\{\eta > 0 : \rho(\mathbb{E}[\boldsymbol{M}_t^+(\eta) \otimes \boldsymbol{M}_t^+(\eta)]) < 1\}.$$

In words, $\eta_\star$ is the largest LR below which every perturbation in the convex subspace decays in mean square, and Corollary D.8 says that above it some perturbation grows without bound. This is the positive-curvature failure of Section 3. The stability notion in Definition B.2 is the bounded affine notion of the main text. The retained subspace still differs from Definition 3.3 because $\Pi_{t_0}^{>0}$ drops the zero-curvature directions that the main text's $\Pi_{t_0}^{+}$ keeps. We project after every update, so a batch Hessian cannot move mass from the discarded directions back into the retained subspace. If the nonnegative-curvature subspace contains no zero eigenmode, the two projections coincide and the two critical LRs are equal. Proposition B.3 justifies computing this threshold from strict homogeneous stability. Below the threshold, strict decay implies bounded affine moments. Above it, differences of affine trajectories driven by the same batches rule out boundedness for every start. The proof is in Appendix E.6. The two stable sets can differ at the endpoint, but their suprema agree.

Zero modes can invalidate this equality. For example, take SGD with $\boldsymbol{H}_{\mathcal{B}_t} = \mathrm{diag}(1, 0)$ and residual gradient $\boldsymbol{r}_t = (0, \xi_t)^\top$, where the ξ_t are independent uniform signs. From zero initial error, the second coordinate satisfies

$$\varepsilon_{t,2} = -\eta\sum_{s=0}^{t-1}\xi_s, \qquad \mathbb{E}[\varepsilon_{t,2}^2] = t\eta^2.$$

Thus no positive LR is stable on the nonnegative subspace. The strictly positive homogeneous system retains only the first coordinate and has critical LR 2. The threshold statement above concerns this strictly positive projection and does not assert equality with the main-text threshold when zero modes remain.

Why Mean Square Stability. For a stochastic system, the mean trajectory does not describe the size of a typical perturbation. Positive and negative deviations can cancel in expectation while their squared norm grows. We therefore study $\mathbb{E}[\|\boldsymbol{\epsilon}_t\|^2]$. This is a standard stability measure in many analyses of stochastic linear systems induced by SGD (Wu et al., 2018a; Ma & Ying, 2021; Wu et al., 2022b; Mulayoff & Michaeli, 2024). Our definition requires this quantity to remain bounded over time, allowing persistent fluctuations without requiring individual trajectories to converge.

Mean-square and almost-sure stability. Consider the homogeneous recursion $\boldsymbol{\epsilon}_{t+1} = \boldsymbol{M}_t\boldsymbol{\epsilon}_t$ with i.i.d. $\boldsymbol{M}_t$ of finite second moment. Recall the strict notion of Appendix D.1. The recursion is strictly mean-square stable if $\mathbb{E}[\|\boldsymbol{\epsilon}_t\|^2] \to 0$ for every deterministic initial state, and this holds if and only if the second-moment map below has spectral radius less than one. The main-text definition (Definition 3.1) asks only for a bounded second moment. The AGM critical LR in Definition B.2 uses this same notion. Proposition B.3 shows that its threshold agrees with the strict homogeneous threshold on the same positive-mean subspace. Boundedness at the endpoint may depend on affine forcing. The spectral analysis below and in Appendix G uses strict homogeneous stability. Exponential mean square decay requires constants $K < \infty$ and $0 < \varrho < 1$ such that

$$\mathbb{E}[\|\boldsymbol{\epsilon}_t\|^2] \le K\varrho^t\|\boldsymbol{\epsilon}_0\|^2.$$

Since $\boldsymbol{M}_t$ is independent of $\boldsymbol{\epsilon}_t$, the second moment $\mathbb{E}[\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\top]$ evolves as $\mathbb{E}[\boldsymbol{\epsilon}_{t+1}\boldsymbol{\epsilon}_{t+1}^\top] = \mathcal{T}(\mathbb{E}[\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\top])$, where

$$\mathcal{T}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{M}_t\boldsymbol{X}\boldsymbol{M}_t^\top].$$

Its matrix representation is the $\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]$ of Proposition D.3, and strict mean-square stability is the spectral condition $\rho(\mathcal{T}) < 1$. This characterization makes mean-square stability useful for analyzing SGD (Wu et al., 2018a; Ma & Ying, 2021; Wu et al., 2022b; Mulayoff & Michaeli, 2024). Almost-sure stability requires $\|\boldsymbol{\epsilon}_t\| \to 0$ with probability one. Its exponential version is governed by the top Lyapunov exponent of the random matrix product,

$$\limsup_{t\to\infty} \frac{1}{t}\log\|\boldsymbol{M}_{t-1}\cdots\boldsymbol{M}_0\| < 0 \qquad \text{almost surely,}$$

under the usual integrability conditions (Furstenberg & Kesten, 1960; Bougerol et al., 2012).

The two notions are not equivalent. Exponential mean square decay implies almost-sure exponential stability by Markov's inequality and the Borel–Cantelli lemma. The converse is false because rare expanding trajectories can dominate the second moment without changing the behavior of a typical sample path. For example, consider the scalar recursion $\varepsilon_{t+1} = a_t\varepsilon_t$, where the a_t are i.i.d. and

$$a_t = \begin{cases} 1/2, & \text{with probability } 0.99, \\ 10, & \text{with probability } 0.01. \end{cases}$$

Here

$$\mathbb{E}[\log a_t] = 0.99\log\tfrac{1}{2} + 0.01\log 10 \approx -0.66 < 0,$$
$$\mathbb{E}[a_t^2] = 0.99\cdot\tfrac{1}{4} + 0.01\cdot 100 = 1.2475 > 1.$$

Since $\log|\varepsilon_t| = \log|\varepsilon_0| + \sum_{s=0}^{t-1}\log a_s$, the strong law of large numbers gives $\frac{1}{t}\log|\varepsilon_t| \to \mathbb{E}[\log a_t] < 0$ almost surely, so $|\varepsilon_t| \to 0$ almost surely, whereas $\mathbb{E}[|\varepsilon_t|^2] = \mathbb{E}[a_t^2]^t|\varepsilon_0|^2$ by independence, so $\mathbb{E}[|\varepsilon_t|^2]$ grows exponentially. The second moment can therefore reveal growth that an almost-sure criterion does not detect. We use mean-square stability because it controls the expected perturbation energy and admits a closed second-moment recursion. The experiments use a single-trajectory diagnostic, whose threshold need not coincide with this population threshold.

D.3 Stochastic Power Iteration for EoC

We state the accuracy of the stochastic power iteration of Appendix B.2. The theorem separates the error due to the finite horizon, the curvature noise, and the bisection resolution.

The system. We use the effective curvature of Appendix D.2, writing the population preconditioned Hessian as $\boldsymbol{D}_{t_0}$, the $\boldsymbol{D}$ of that section:

$$\boldsymbol{A} \coloneqq \boldsymbol{D}_{t_0} + \lambda_{\mathrm{WD}}\boldsymbol{I}, \qquad \boldsymbol{A}_{\mathcal{B}_t} \coloneqq \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2} + \lambda_{\mathrm{WD}}\boldsymbol{I},$$
$$\boldsymbol{\Delta}_t \coloneqq \boldsymbol{A}_{\mathcal{B}_t} - \boldsymbol{A}.$$

Thus the deterministic weight-decay term is included in both curvatures and cancels from the noise $\boldsymbol{\Delta}_t$. For SGD without weight decay, $\boldsymbol{A} = \boldsymbol{H}_{t_0}$, so the same procedure estimates the minimum eigenvalue of the raw population Hessian. Let $\boldsymbol{u}_{\min}$ be a unit eigenvector of $\boldsymbol{A}$ for its smallest eigenvalue. The shift to estimate is

$$\lambda_\star \coloneqq \max\{0, -\lambda_{\min}(\boldsymbol{A})\}, \tag{39}$$

the smallest nonnegative shift making $\boldsymbol{A} + \lambda\boldsymbol{I}$ positive semidefinite.

A query at shift $\lambda \in [0, \Lambda]$ starts from the same fixed unit vector $\boldsymbol{x}_0$ and uses the inner stepsize η_{p} for T_{p} steps:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \big[\boldsymbol{I} - \eta_{\mathrm{p}}(\boldsymbol{A}_{\mathcal{B}_t} + \lambda\boldsymbol{I})\big]\boldsymbol{x}_t \\ &= \big[\boldsymbol{M}(\lambda) - \eta_{\mathrm{p}}\boldsymbol{\Delta}_t\big]\boldsymbol{x}_t, \qquad t = 0, \dots, T_{\mathrm{p}} - 1, \\ \boldsymbol{M}(\lambda) &\coloneqq \boldsymbol{I} - \eta_{\mathrm{p}}(\boldsymbol{A} + \lambda\boldsymbol{I}). \end{aligned} \tag{40}$$

Here η_{p} is the power-iteration stepsize, distinct from the training LR η, and T_{p} is its horizon, distinct from the local-simulation horizon T. A query explodes if $\max_{0\le t\le T_{\mathrm{p}}} \|\boldsymbol{x}_t\| > B$, where $B > 1$ is the cutoff. Starting from $[0, \Lambda]$, Algorithm 1 queries the midpoint of the current bracket. An exploding query replaces the lower endpoint and a non-exploding query replaces the upper endpoint. After N queries the output $\widehat{\lambda}$ is the upper endpoint, and $\widehat{\lambda}_{\min} = -\widehat{\lambda}$.

Within each query, let $\mathcal{F}_t$ contain all observations of earlier queries and the current observations $\boldsymbol{\Delta}_0, \dots, \boldsymbol{\Delta}_{t-1}$. The queried shift and $\boldsymbol{x}_t$ are therefore $\mathcal{F}_t$-measurable. All vector norms below are Euclidean and matrix norms are spectral.

Assumption D.9 (Bounded curvature noise)**.** For each query and each step t, the symmetric noise $\boldsymbol{\Delta}_t$ is $\mathcal{F}_{t+1}$-measurable and satisfies

$$\mathbb{E}[\boldsymbol{\Delta}_t \mid \mathcal{F}_t] = \boldsymbol{0}, \qquad \|\boldsymbol{\Delta}_t\| \le \sigma \quad \text{almost surely}, \tag{41}$$

for a constant $\sigma > 0$.

Fresh independent mini-batches, as in Appendix D.1, give the conditional mean-zero property. The almost-sure norm bound is an additional assumption used by the concentration inequalities in the proof.

Assumption D.10 (Bracket and initial overlap)**.** The population curvature satisfies $\lambda_{\min}(\boldsymbol{A}) < 0$ and $\lambda_\star \le \Lambda$. The fixed initialization satisfies

$$\|\boldsymbol{x}_0\| = 1, \qquad |\boldsymbol{u}_{\min}^\top \boldsymbol{x}_0| \ge \varepsilon, \qquad 0 < \varepsilon \le 1.$$

The bracket must contain the target. The overlap supplies a nonzero initial component in the direction that grows below $\lambda_\star$. Initializing with the checkpoint momentum does not by itself ensure this condition.

Assumption D.11 (Power-iteration stepsize)**.** The inner stepsize satisfies

$$\eta_{\mathrm{p}} > 0, \qquad \eta_{\mathrm{p}}\big(\lambda_{\max}(\boldsymbol{A}) + \Lambda\big) \le 1. \tag{42}$$

This makes the mean transition $\boldsymbol{M}(\lambda)$ positive semidefinite for every queried shift. Above $\lambda_\star$ its norm is $1 - \eta_{\mathrm{p}}(\lambda - \lambda_\star)$, which controls the probability of a false explosion. The explosion argument below $\lambda_\star$ does not need this stepsize condition.

The following result applies to any fixed finite horizon. It imposes no upper bound on that horizon.

Theorem D.12 (Accuracy of the stochastic power iteration)**.** *Let Assumptions D.9 to D.11 hold. Fix a cutoff $B > 1$, positive integers N and T_{p}, and $0 < \delta < 1$. Define*

$$g_- \coloneqq \frac{4\eta_{\mathrm{p}}\sigma^2 B^2}{\varepsilon^2}\log\frac{N}{\delta}, \qquad g_+ \coloneqq 2\eta_{\mathrm{p}}\sigma^2 \max\left\{1, \frac{\log(N/\delta)}{\log B}\right\},$$

$$e_{\mathrm{p}} := \frac{(2B/\varepsilon)^{1/T_{\mathrm{p}}} - 1}{\eta_{\mathrm{p}}}. \tag{43}$$

Then the bisection output satisfies, with probability at least $1 - \delta$*,*

$$\lambda_\star - \max\{g_-, e_{\mathrm{p}}\} \le \widehat{\lambda} \le \lambda_\star + g_+ + \frac{\Lambda}{2^N}. \tag{44}$$

The finite-horizon term obeys $e_{\mathrm{p}} \le g_-$ *if and only if*

$$T_{\mathrm{p}} \ge \left\lceil \frac{\log(2B/\varepsilon)}{\log(1 + \eta_{\mathrm{p}} g_-)} \right\rceil. \tag{45}$$

Reading the bound. The lower error is the larger of the noise term g_- and the finite-horizon term e_{p}. For fixed B, ε, and η_{p},

$$e_{\mathrm{p}} = \frac{\log(2B/\varepsilon)}{\eta_{\mathrm{p}} T_{\mathrm{p}}} + O(T_{\mathrm{p}}^{-2}) \qquad \text{as } T_{\mathrm{p}} \to \infty.$$

This is the margin needed for the initial component of size ε to grow past B within the allotted steps. Above $\lambda_\star$, the error consists of the noise term g_+ and the bisection resolution $\Lambda 2^{-N}$. The displayed noise terms do not decrease with the horizon when the stepsize, cutoff, overlap, confidence, and number of queries are fixed. A longer horizon therefore reduces the finite-horizon term but does not make this bound vanish at a fixed stepsize. The theorem assumes the bracket and overlap and does not establish them. It also assumes that the population minimum is negative. It bounds the estimation error under that assumption and does not certify the sign of the population minimum from the sign of the returned estimate. Its probability guarantee is for each fixed choice of T_{p}. The proof is in Appendix F.

D.4 Compensated Local Simulation for EoS

The detector tests a compensated local model, whereas EoS is defined by the positive eigenspace of the uncompensated effective curvature. The theorem below compares its population output with that projected critical LR. It separates the change caused by compensation from the error caused by a finite simulation horizon.

Curvature at the checkpoint. We continue with the checkpoint, frozen preconditioner and local model of Appendix D.1. Recall the effective curvature and its mini-batch version from Appendix D.3,

$$\boldsymbol{A} \coloneqq \boldsymbol{D}_{t_0} + \lambda_{\mathrm{WD}}\boldsymbol{I}, \qquad \boldsymbol{A}_{\mathcal{B}_t} \coloneqq \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2} + \lambda_{\mathrm{WD}}\boldsymbol{I}, \qquad \boldsymbol{D}_{t_0} \coloneqq \boldsymbol{P}^{-1/2}\boldsymbol{H}_{t_0}\boldsymbol{P}^{-1/2}, \tag{46}$$

and $\boldsymbol{\Delta}_t \coloneqq \boldsymbol{A}_{\mathcal{B}_t} - \boldsymbol{A}$ is the deviation of the mini-batch effective curvature from its mean. It is symmetric with $\mathbb{E}\boldsymbol{\Delta}_t = 0$. Let Π^+ be the orthogonal projector onto the strictly positive eigenspace of $\boldsymbol{A}$, of dimension $p \geq 1$, and choose orthogonal coordinates adapted to it, so that

$$\boldsymbol{A} = \mathrm{diag}(\boldsymbol{A}^+, \boldsymbol{A}^-), \qquad \boldsymbol{A}^+ \succ 0, \qquad \boldsymbol{A}^- \preceq 0,$$

where $\boldsymbol{A}^+$ is the retained $p \times p$ block and $\boldsymbol{A}^-$ the block on the discarded directions, absent when $p = d$. For a symmetric $d \times d$ matrix $\boldsymbol{M}$ written in these coordinates, $\boldsymbol{M}^+$ denotes its retained $p \times p$ block and $\boldsymbol{M}^-$ its discarded block, so that $\Pi^+\boldsymbol{M}\Pi^+ = \mathrm{diag}(\boldsymbol{M}^+, 0)$ is $\boldsymbol{M}^+$ padded with zeros. In particular $\boldsymbol{\Delta}_t^+$ is the retained block of the noise, of the same size as $\boldsymbol{A}^+$, and $\boldsymbol{\Delta}_t^\perp \coloneqq \boldsymbol{\Delta}_t - \Pi^+\boldsymbol{\Delta}_t\Pi^+$ collects the discarded and cross blocks. We use $L \coloneqq \lambda_{\max}(\boldsymbol{A}) = \lambda_{\max}(\boldsymbol{A}^+)$ throughout. Thus the compensated mean has largest eigenvalue $L + \lambda_{\mathsf{P}}$. The compensation shifts the effective curvature to $\boldsymbol{A}_{\mathsf{P}} \coloneqq \boldsymbol{A} + \lambda_{\mathsf{P}}\boldsymbol{I}$.

The compensated local simulation. The detector runs the local AGM update of Appendix D.1 on the compensated loss G_t. We retain $\boldsymbol{\epsilon}_t$ for the displacement from the checkpoint, so the simulated parameter is $\boldsymbol{w}_{t_0} + \boldsymbol{\epsilon}_t$. A run at LR $\eta > 0$ from a prescribed displacement $\boldsymbol{\epsilon}_0$ and the checkpoint momentum $\boldsymbol{m}_0 = \boldsymbol{m}_{t_0}$ follows

$$\begin{aligned} \boldsymbol{m}_{t+1} &= \beta_1\boldsymbol{m}_t + (1-\beta_1)\big[\boldsymbol{g}_{\mathcal{B}_t} + (\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{P})\boldsymbol{\epsilon}_t\big], \\ \boldsymbol{\epsilon}_{t+1} &= (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{\epsilon}_t - \eta\boldsymbol{P}^{-1}\boldsymbol{m}_{t+1} - \eta\lambda_{\mathrm{WD}}\boldsymbol{w}_{t_0}. \end{aligned} \tag{47}$$

The two critical LRs. The homogeneous differences of two runs driven by the same mini-batches satisfy a momentum recurrence. In preconditioned position coordinates this recurrence is

$$\boldsymbol{x}_{t+1} = \big[(1+b)\boldsymbol{I} - \alpha(\boldsymbol{C} + \boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t - b\boldsymbol{x}_{t-1}, \qquad b \coloneqq \beta_1(1-\eta\lambda_{\mathrm{WD}}), \quad \alpha \coloneqq \eta(1-\beta_1).$$

For the compensated system, the mean and noise are $(\boldsymbol{C}, \boldsymbol{\Delta}_t) = (\boldsymbol{A}_{\mathsf{P}}, \boldsymbol{\Delta}_t)$. For the projected system, they are $(\boldsymbol{A}^+, \boldsymbol{\Delta}_t^+)$, on the retained p-dimensional subspace. The derivation and relation to the augmented state of Appendix D.1 are given in Appendix G.1.

Definition D.13 (Critical LRs)**.** For either homogeneous system, strict mean-square stability means that $\mathbb{E}\|\boldsymbol{x}_t\|^2 + \mathbb{E}\|\boldsymbol{x}_{t-1}\|^2 \to 0$ for every deterministic initial pair. The projected critical LR $\eta_\star$ and the compensated critical LR $\eta_\star^{\mathsf{P}}$ are the respective suprema of the strictly stable LRs.

By Proposition B.3, $\eta_\star$ is the projected critical LR defined by bounded affine moments in Definition B.2. The same argument identifies $\eta_\star^{\mathsf{P}}$ with the bounded affine threshold of the compensated system when $\boldsymbol{A}_{\mathsf{P}} \succ 0$. By Corollary G.4, the strictly stable rates form the intervals $(0, \eta_\star)$ and $(0, \eta_\star^{\mathsf{P}})$, respectively. Their second-moment maps are denoted $\mathcal{T}_\eta$ and $\mathcal{T}_\eta^{\mathsf{P}}$, as defined in (62).

The detector. Fix a search window $\mathcal{I} = [\eta - w, \eta + w] \subset (0, \infty)$ around the training LR η, an integer horizon $T \geq 1$ and a cutoff $M > 1$. When $\lambda_{\mathrm{WD}} > 0$, assume $\mathcal{I} \subset (0, 1/\lambda_{\mathrm{WD}}]$, so each candidate $\eta' \in \mathcal{I}$ has effective momentum $b = \beta_1(1-\eta'\lambda_{\mathrm{WD}}) \in [0, \beta_1]$. The population version of Algorithm 3 returns

$$\widehat{\eta}_\star(M, T) \coloneqq \sup\Big\{\eta' \in \mathcal{I} : \max_{0 \leq t \leq T} \mathbb{E}\,\|\boldsymbol{\epsilon}_t(\eta')\|^2 \leq M^2 \text{ for every } \|\boldsymbol{\epsilon}_0\| = 1\Big\}, \tag{48}$$

where $\boldsymbol{\epsilon}_t(\eta')$ follows (47) with prescribed momentum $\boldsymbol{m}_0 = \boldsymbol{m}_{t_0}$, and the expectation is over mini-batches. Differences between this population detector and its implementation are listed in Appendix B.3.

Assumption D.14 (Frozen compensated model)**.** Assumption D.1 holds. Fix $0 \le \beta_1 < 1$, $\lambda_{\mathrm{WD}} \ge 0$ and $\lambda_{\mathsf{P}} > 0$. The compensation satisfies $\boldsymbol{A}_{\mathsf{P}} = \boldsymbol{A} + \lambda_{\mathsf{P}} \boldsymbol{I} \succ 0$. All parameters and the batch law remain fixed while η varies.

The preconditioner and batch law are frozen at the checkpoint. The Hessian and gradient from the same mini-batch may be correlated, and no small-noise assumption is imposed. Positive definiteness of $\boldsymbol{A}_{\mathsf{P}}$ requires compensation to remove all negative and zero curvature directions. If the compensation is estimated from data, the theorem applies after fixing that estimate only when the resulting mean is positive definite and the simulation batches still satisfy Assumption D.1. In particular, after the checkpoint and compensation are fixed, those batches must remain i.i.d. and independent of the prescribed initial state. The estimated shift alone does not verify positive definiteness.

Assumption D.15 (Acceptance at the compensated critical LR)**.** The compensated critical LR $\eta_\star^{\mathsf{P}}$ belongs to the window $\mathcal{I}$ and is accepted by the detector (48): $\max_{0 \le t \le T} \mathbb{E} \left\| \boldsymbol{\epsilon}_t(\eta_\star^{\mathsf{P}}) \right\|^2 \le M^2$ for every $\|\boldsymbol{\epsilon}_0\| = 1$, with the prescribed initial momentum.

The boundary condition $\rho(\mathcal{T}_{\eta_\star^{\mathsf{P}}}^{\mathsf{P}}) = 1$ does not imply acceptance. Transient growth, checkpoint momentum, and affine forcing can exceed the cutoff. Proposition G.11 gives an explicit sufficient condition on M. Acceptance of $\eta_\star^{\mathsf{P}}$ is used for the upper bound below. The lower bound only needs $\eta_\star^{\mathsf{P}} \in \mathcal{I}$ and a nonempty accepted set.

The quantity that determines stability. Unqualified vector norms are Euclidean, matrix norms are spectral, and $\|\cdot\|_F$ is the Frobenius norm. For a positive definite mean $\boldsymbol{C}$ with its corresponding centered noise, and $-1 < b < 1$, let

$$c_b := \frac{1-b}{1+b}, \qquad k_b := \frac{b}{1-b^2}.$$

For a symmetric test matrix, $[\boldsymbol{C}, \boldsymbol{X}] := \boldsymbol{C}\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C}$ is the commutator. Define

$$q_{\boldsymbol{C},b}(\boldsymbol{X}) := \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) + c_b\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) - k_b\,\|[\boldsymbol{C},\boldsymbol{X}]\|_F^2}{2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)}, \qquad 0 \ne \boldsymbol{X} \succeq 0, \tag{49}$$

$$r_{\boldsymbol{C}}(b) := \max_{\boldsymbol{X} \succeq 0,\ \|\boldsymbol{X}\|_F = 1} q_{\boldsymbol{C},b}(\boldsymbol{X}). \tag{50}$$

The noise is $\boldsymbol{\Delta}_t^+$ when $\boldsymbol{C} = \boldsymbol{A}^+$ and $\boldsymbol{\Delta}_t$ when $\boldsymbol{C} = \boldsymbol{A}_{\mathsf{P}}$. The maximum is attained on the compact set of positive semidefinite matrices of unit Frobenius norm. The exact criterion in Theorem G.1 shows that the corresponding recurrence is strictly mean-square stable precisely when $(1-b)/\alpha > r_{\boldsymbol{C}}(b)$. Its maximizers determine the curvatures in the following bound.

Theorem D.16 (Finite-horizon accuracy of the population EoS detector)**.** *Under Assumptions D.14 and D.15,*

$$\min\{\Phi(\mu_\star), 0\} - H_T \;\le\; \frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star} \;\le\; \max\{\Phi(L) + \delta_{\mathrm{cov}}, 0\}, \tag{51}$$

where $L = \lambda_{\max}(\boldsymbol{A})$ as fixed above, and the terms are defined as follows.

(i) Critical curvatures. *Set $b_\star := \beta_1(1 - \eta_\star \lambda_{\mathrm{WD}})$ and $b_\star^{\mathsf{P}} := \beta_1(1 - \eta_\star^{\mathsf{P}} \lambda_{\mathrm{WD}})$. Choose*

$$\begin{aligned} \boldsymbol{X}_\star &\in \operatorname*{arg\,max}_{\boldsymbol{X}\succeq 0,\ \|\boldsymbol{X}\|_F=1} q_{\boldsymbol{A}^+, b_\star}(\boldsymbol{X}), \quad \mu_\star := \operatorname{tr}(\boldsymbol{A}^+ \boldsymbol{X}_\star^2), \\ \boldsymbol{X}_\star^{\mathsf{P}} &\in \operatorname*{arg\,max}_{\boldsymbol{X}\succeq 0,\ \|\boldsymbol{X}\|_F=1} q_{\boldsymbol{A}_{\mathsf{P}}, b_\star^{\mathsf{P}}}(\boldsymbol{X}), \quad \mu_\star^{\mathsf{P}} := \operatorname{tr}\big(\boldsymbol{A}_{\mathsf{P}} (\boldsymbol{X}_\star^{\mathsf{P}})^2\big). \end{aligned} \tag{52}$$

The first maximization is over $p \times p$ matrices and the second over $d \times d$ matrices. Then

$$\lambda_{\min}(\boldsymbol{A}^+) \le \mu_\star \le L, \qquad \lambda_{\min}(\boldsymbol{A}_{\mathsf{P}}) \le \mu_\star^{\mathsf{P}} \le L + \lambda_{\mathsf{P}}.$$

(ii) Covariance mismatch. *With $\boldsymbol{X}^+$ the retained block of $\boldsymbol{X}$,*

$$\delta_{\mathrm{cov}} := \max_{\boldsymbol{X}\succeq 0,\ \|\boldsymbol{X}\|_F=1} \frac{\mathbb{E}\operatorname{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t \boldsymbol{X}\boldsymbol{\Delta}_t) - \mathbb{E}\operatorname{tr}(\boldsymbol{X}^+\boldsymbol{\Delta}_t^+ \boldsymbol{X}^+\boldsymbol{\Delta}_t^+)}{2\operatorname{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)} \ge 0. \tag{53}$$

(iii) Compensation. *With $c_\star := (1 - b_\star)/(1 + b_\star)$, define, for $0 \le s \le L$,*

$$\Phi(s) := \frac{c_\star \lambda_{\mathsf{P}}}{2} - \frac{\lambda_{\mathsf{P}}}{s + \lambda_{\mathsf{P}}}\Big(r_{\boldsymbol{A}^+}(b_\star) - \frac{c_\star s}{2}\Big). \tag{54}$$

This function is increasing and satisfies $\Phi(L) \le c_\star \lambda_{\mathsf{P}}/2$.

(iv) Finite horizon. *The recurrence coefficient at the compensated critical rate is $\alpha_\star^{\mathsf{P}} := (1 - \beta_1)\eta_\star^{\mathsf{P}}$. Every accepted LR has moment spectral radius at most*

$$\rho_T := \left[\frac{M(1+\beta_1)}{1-\beta_1}\right]^{2/T}. \tag{55}$$

The finite-horizon term is

$$H_T := (\rho_T - 1)\left[\frac{2b_\star^{\mathsf{P}}}{(\rho_T + b_\star^{\mathsf{P}})\alpha_\star^{\mathsf{P}}} + \frac{(1 - b_\star^{\mathsf{P}}/\rho_T)^2}{2\mu_\star^{\mathsf{P}}(\alpha_\star^{\mathsf{P}})^2}\right]. \tag{56}$$

If $\beta_1 \lambda_{\mathrm{WD}} = 0$, the clipping at zero can be omitted from (51). For $\beta_1 = 0$, the same formula gives $H_T = (\rho_T - 1)/(2\mu_\star^{\mathsf{P}}(\eta_\star^{\mathsf{P}})^2)$. For fixed model parameters, $H_T = O((1 + \log M)/T)$ as $(1 + \log M)/T \to 0$.

Reading the bound. The compensation and finite-horizon errors have different roles.

- *Compensation changes the critical rate.* The curvature $\mu_\star$ need not be the largest eigenvalue L. Noise on weakly curved retained directions can make $\mu_\star$ much smaller, and then the lower bound can be weaker. The term δ_{cov} measures the additional noise seen outside the retained block. A test matrix supported on the retained subspace gives zero mismatch, so the maximum is nonnegative. It is finite, vanishes when the discarded and cross noise blocks vanish, and is not assumed small. Lemma G.7 bounds it using their second moments.
- *A finite horizon can accept a LR above the compensated critical rate.* Acceptance of $\eta_\star^{\mathsf{P}}$ ensures $\widehat{\eta}_\star(M,T) \ge \eta_\star^{\mathsf{P}}$, so this error lies in $[-H_T, 0]$. As $(1 + \log M)/T \to 0$, $\rho_T \to 1$ and $H_T \to 0$. The compensation error remains. Proposition G.11 gives a sufficient cutoff for accepting $\eta_\star^{\mathsf{P}}$ and shows that a polynomially growing cutoff is possible.

Weight decay changes the effective momentum when the LR changes. The clipping in (51) accounts for this change. Without weight decay, or without momentum, the comparison transfers directly and no clipping is needed.

The full proof is in Appendix G. Corollary G.8 gives the simpler compensation and noise bound used in Theorem B.5, with a common second-moment noise level v. The expectation and all-unit-start requirements belong to the population detector. The practical differences are described in Appendix B.3.

E PROOFS FOR THE LOCAL LINEAR SYSTEM AND THE STABILITY RESULTS

This appendix proves the statements of Appendices D.1 and D.2 in the order they are stated. Appendix E.1 derives the augmented linear system (Proposition D.2) and its compensated form. Appendix E.2 derives the moment recursion of a stochastic linear system (Proposition D.3). Appendix E.3 derives the moment recursion for SGD (Corollary D.4) from the two propositions. Appendix E.4 proves the two negative-curvature results (Lemma D.5 and Corollary D.6) and Appendix E.5 the divergence above the critical LR (Corollary D.8). Theorem D.7 is restated from Corollary G.4 and proved in Appendix G.5.3. Appendix E.6 proves that bounded affine stability gives the same critical LR as the spectral criterion (Proposition B.3). Throughout, ρ denotes the spectral radius and $\|\cdot\|$ the spectral norm.

E.1 PROOF OF PROPOSITION D.2

Proof of Proposition D.2. The idea is to center the displacement at $\boldsymbol{x}_\star$, multiply each of the two updates in (31) by the power of $\boldsymbol{P}$ that appears in the corresponding block of $\boldsymbol{z}_t$, and identify the blocks of $\boldsymbol{M}_t(\eta)$ and $\boldsymbol{q}_t$. Here $\boldsymbol{x}_\star$ is the fixed reference displacement in the proposition and need not be a stationary point. Write $\bar{\boldsymbol{\epsilon}}_t := \boldsymbol{\epsilon}_t - \boldsymbol{x}_\star$ for the centered displacement, so that $\boldsymbol{z}_t = (\eta \boldsymbol{P}^{-1/2}\boldsymbol{m}_t, \boldsymbol{P}^{1/2}\bar{\boldsymbol{\epsilon}}_t)$. Recall the forcing in the momentum block,

$$\boldsymbol{h}_t = (1-\beta_1)\eta \boldsymbol{P}^{-1/2}(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{x}_\star).$$

Recall that $\boldsymbol{D}_{\mathcal{B}_t} = \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2}$ is the preconditioned mini-batch Hessian, as defined before Proposition D.2.

Step 1. The momentum block. Multiplying the momentum update in (31) by $\eta\boldsymbol{P}^{-1/2}$ and substituting $\boldsymbol{\epsilon}_t = \bar{\boldsymbol{\epsilon}}_t + \boldsymbol{x}_\star$ gives the first block of $\boldsymbol{z}_{t+1}$,

$$\begin{aligned}
\eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_{t+1} &= \beta_1\,\eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_t + (1-\beta_1)\eta\boldsymbol{P}^{-1/2}(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon}_t) && \text{momentum update}\\
&= \beta_1\,\eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_t + (1-\beta_1)\eta\boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\bar{\boldsymbol{\epsilon}}_t \\
&\quad + (1-\beta_1)\eta\boldsymbol{P}^{-1/2}(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{x}_\star) && \boldsymbol{\epsilon}_t = \bar{\boldsymbol{\epsilon}}_t + \boldsymbol{x}_\star\\
&= \beta_1\,\eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_t + (1-\beta_1)\eta\,\boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2}\,\boldsymbol{P}^{1/2}\bar{\boldsymbol{\epsilon}}_t \\
&\quad + (1-\beta_1)\eta\boldsymbol{P}^{-1/2}(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{x}_\star) && \boldsymbol{P}^{-1/2}\boldsymbol{P}^{1/2} = \boldsymbol{I}\\
&= \beta_1\,\eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_t + (1-\beta_1)\eta\boldsymbol{D}_{\mathcal{B}_t}\,\boldsymbol{P}^{1/2}\bar{\boldsymbol{\epsilon}}_t + \boldsymbol{h}_t. && \text{definitions of } \boldsymbol{D}_{\mathcal{B}_t}, \boldsymbol{h}_t
\end{aligned}$$

Step 2. The displacement block. Subtracting $\boldsymbol{x}_\star$ from both sides of the displacement update and multiplying by $\boldsymbol{P}^{1/2}$ gives the second block,

$$\begin{aligned}
\boldsymbol{P}^{1/2}\bar{\boldsymbol{\epsilon}}_{t+1} &= \boldsymbol{P}^{1/2}\bar{\boldsymbol{\epsilon}}_t - \eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_{t+1} - \eta\lambda_{\mathrm{WD}}\boldsymbol{P}^{1/2}(\bar{\boldsymbol{\epsilon}}_t + \boldsymbol{x}_\star + \boldsymbol{w}_{t_0}) && \boldsymbol{\epsilon}_t = \bar{\boldsymbol{\epsilon}}_t + \boldsymbol{x}_\star\\
&= (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{P}^{1/2}\bar{\boldsymbol{\epsilon}}_t - \eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_{t+1} - \eta\lambda_{\mathrm{WD}}\boldsymbol{P}^{1/2}(\boldsymbol{x}_\star + \boldsymbol{w}_{t_0}) && \text{collect } \bar{\boldsymbol{\epsilon}}_t \text{ terms}\\
&= -\beta_1\,\eta\boldsymbol{P}^{-1/2}\boldsymbol{m}_t + \big((1-\eta\lambda_{\mathrm{WD}})\boldsymbol{I} - (1-\beta_1)\eta\boldsymbol{D}_{\mathcal{B}_t}\big)\boldsymbol{P}^{1/2}\bar{\boldsymbol{\epsilon}}_t \\
&\quad - \boldsymbol{h}_t - \eta\lambda_{\mathrm{WD}}\boldsymbol{P}^{1/2}(\boldsymbol{x}_\star + \boldsymbol{w}_{t_0}). && \text{insert Step 1}
\end{aligned}$$

In the state $\boldsymbol{z}_t$, the two identities read $\boldsymbol{z}_{t+1} = \boldsymbol{M}_t(\eta)\boldsymbol{z}_t + \boldsymbol{q}_t$ with $\boldsymbol{M}_t(\eta)$ and $\boldsymbol{q}_t$ as in (32).

Step 3. Compensation. The gradient of the compensated loss G_t of Appendix D.1 is

$$\begin{aligned}
\nabla G_t(\boldsymbol{\epsilon}) &= \nabla\mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0}) + \nabla^2\mathcal{L}_{\mathcal{B}_t}(\boldsymbol{w}_{t_0})\boldsymbol{\epsilon} + \lambda_{\mathsf{P}}\boldsymbol{P}\boldsymbol{\epsilon} && \nabla\tfrac{\lambda_{\mathsf{P}}}{2}\boldsymbol{\epsilon}^\top\boldsymbol{P}\boldsymbol{\epsilon} = \lambda_{\mathsf{P}}\boldsymbol{P}\boldsymbol{\epsilon}, \text{ as } \boldsymbol{P} \text{ is symmetric}\\
&= \boldsymbol{g}_{\mathcal{B}_t} + (\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{P})\boldsymbol{\epsilon}, && \text{definitions of } \boldsymbol{g}_{\mathcal{B}_t} \text{ and } \boldsymbol{H}_{\mathcal{B}_t}
\end{aligned}$$

so the penalty replaces the mini-batch Hessian $\boldsymbol{H}_{\mathcal{B}_t}$ by $\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{P}$ in the momentum update. In preconditioned coordinates,

$$\begin{aligned}
\boldsymbol{P}^{-1/2}(\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{P})\boldsymbol{P}^{-1/2} &= \boldsymbol{P}^{-1/2}\boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{P}^{-1/2} + \lambda_{\mathsf{P}}\,\boldsymbol{P}^{-1/2}\boldsymbol{P}\boldsymbol{P}^{-1/2}\\
&= \boldsymbol{D}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I}. && \boldsymbol{P}^{-1/2}\boldsymbol{P}\boldsymbol{P}^{-1/2} = \boldsymbol{I}
\end{aligned}$$

Repeating Steps 1 and 2 with $\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathrm{P}}\boldsymbol{P}$ in place of $\boldsymbol{H}_{\mathcal{B}_t}$, and hence $\boldsymbol{D}_{\mathcal{B}_t} + \lambda_{\mathrm{P}}\boldsymbol{I}$ in place of $\boldsymbol{D}_{\mathcal{B}_t}$, gives the augmented system with the transition matrix $\widetilde{\boldsymbol{M}}_t(\eta)$ of (33) and with $\boldsymbol{h}_t$ formed from $\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathrm{P}}\boldsymbol{P}$. The weight decay term is not touched by the penalty, so its coefficient remains λ_{WD}.

Step 4. Independence. Under Assumption D.1, $(\boldsymbol{M}_t(\eta), \boldsymbol{q}_t)$ is a function of $\mathcal{B}_t$ alone, while $\boldsymbol{z}_t$ depends only on the initial state and on $\mathcal{B}_0, \ldots, \mathcal{B}_{t-1}$, so $\boldsymbol{M}_t(\eta)$ and $\boldsymbol{q}_t$ are independent of $\boldsymbol{z}_t$. This completes the proof of Proposition D.2. □

E.2 Proof of Proposition D.3

Proof of Proposition D.3. We first derive the closed recursions for the mean and the second moment. A strict spectral bound then makes this finite-dimensional affine recursion converge, including when the current transition and forcing are correlated. Since $\boldsymbol{z}_t$ is a function of $\boldsymbol{z}_0$ and of $(\boldsymbol{M}_s, \boldsymbol{q}_s)_{s<t}$, the pair $(\boldsymbol{M}_t, \boldsymbol{q}_t)$ is independent of $\boldsymbol{z}_t$. Finite second moments propagate by induction:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{z}_{t+1}\|^2 &\le 2\mathbb{E}\|\boldsymbol{M}_t\boldsymbol{z}_t\|^2 + 2\mathbb{E}\|\boldsymbol{q}_t\|^2 && \|\boldsymbol{a}+\boldsymbol{b}\|^2 \le 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2 \\
&\le 2\mathbb{E}\|\boldsymbol{M}_t\|^2\,\mathbb{E}\|\boldsymbol{z}_t\|^2 + 2\mathbb{E}\|\boldsymbol{q}_t\|^2 < \infty. && \text{independence and the induction hypothesis}
\end{aligned}$$

The initial moment is finite by assumption. The Cauchy–Schwarz inequality also gives $\mathbb{E}[\|\boldsymbol{M}_t\|\,\|\boldsymbol{q}_t\|] < \infty$, so the cross terms below are well defined without independence between $\boldsymbol{M}_t$ and $\boldsymbol{q}_t$.

Step 1. The mean. Recall that $\boldsymbol{M} = \mathbb{E}[\boldsymbol{M}_t]$ and $\boldsymbol{z}_t^{(1)} = \mathbb{E}[\boldsymbol{z}_t]$. Then

$$\begin{aligned}
\boldsymbol{z}_{t+1}^{(1)} &= \mathbb{E}[\boldsymbol{M}_t\boldsymbol{z}_t] + \mathbb{E}[\boldsymbol{q}_t] && \text{expectation of } \boldsymbol{z}_{t+1} = \boldsymbol{M}_t\boldsymbol{z}_t + \boldsymbol{q}_t \\
&= \boldsymbol{M}\boldsymbol{z}_t^{(1)} + \mathbb{E}[\boldsymbol{q}_t]. && \boldsymbol{M}_t \text{ is independent of } \boldsymbol{z}_t
\end{aligned}$$

Step 2. The second moment. Recall that $\boldsymbol{z}_t^{(2)} = \mathrm{vec}(\mathbb{E}[\boldsymbol{z}_t\boldsymbol{z}_t^\top])$, $\boldsymbol{S} = \mathbb{E}[\boldsymbol{q}_t \otimes \boldsymbol{M}_t + \boldsymbol{M}_t \otimes \boldsymbol{q}_t]$ and $\boldsymbol{b}_{\mathrm{quad}} = \mathrm{vec}(\mathbb{E}[\boldsymbol{q}_t\boldsymbol{q}_t^\top])$. Expanding the outer product of $\boldsymbol{z}_{t+1} = \boldsymbol{M}_t\boldsymbol{z}_t + \boldsymbol{q}_t$ with itself gives

$$\boldsymbol{z}_{t+1}\boldsymbol{z}_{t+1}^\top = \boldsymbol{M}_t\boldsymbol{z}_t\boldsymbol{z}_t^\top\boldsymbol{M}_t^\top + \boldsymbol{M}_t\boldsymbol{z}_t\boldsymbol{q}_t^\top + \boldsymbol{q}_t\boldsymbol{z}_t^\top\boldsymbol{M}_t^\top + \boldsymbol{q}_t\boldsymbol{q}_t^\top.$$

Vectorize term by term using $\mathrm{vec}(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z}) = (\boldsymbol{Z}^\top \otimes \boldsymbol{X})\mathrm{vec}(\boldsymbol{Y})$. In particular,

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{M}_t\boldsymbol{z}_t\boldsymbol{q}_t^\top) &= (\boldsymbol{q}_t \otimes \boldsymbol{M}_t)\boldsymbol{z}_t, \\
\mathrm{vec}(\boldsymbol{q}_t\boldsymbol{z}_t^\top\boldsymbol{M}_t^\top) &= (\boldsymbol{M}_t \otimes \boldsymbol{q}_t)\boldsymbol{z}_t.
\end{aligned}$$

Taking expectations gives

$$\begin{aligned}
\boldsymbol{z}_{t+1}^{(2)} &= \mathbb{E}\big[(\boldsymbol{M}_t \otimes \boldsymbol{M}_t)\mathrm{vec}(\boldsymbol{z}_t\boldsymbol{z}_t^\top)\big] \\
&\quad + \mathbb{E}\big[(\boldsymbol{q}_t \otimes \boldsymbol{M}_t + \boldsymbol{M}_t \otimes \boldsymbol{q}_t)\boldsymbol{z}_t\big] \\
&\quad + \mathrm{vec}(\mathbb{E}[\boldsymbol{q}_t\boldsymbol{q}_t^\top]) && \text{vectorize term by term} \\
&= \mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]\boldsymbol{z}_t^{(2)} + \boldsymbol{S}\boldsymbol{z}_t^{(1)} + \boldsymbol{b}_{\mathrm{quad}}. && \boldsymbol{z}_t \text{ independent of } (\boldsymbol{M}_t, \boldsymbol{q}_t)
\end{aligned}$$

Combining the two moment recursions gives the block form (34).

Step 3. The spectral radius. For the block lower triangular matrix of the proposition and every scalar z,

$$\begin{aligned}
&\det\begin{bmatrix} z\boldsymbol{I} - \boldsymbol{M} & \boldsymbol{0} \\ -\boldsymbol{S} & z\boldsymbol{I} - \mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t] \end{bmatrix} \\
&\quad = \det(z\boldsymbol{I} - \boldsymbol{M})\,\det(z\boldsymbol{I} - \mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]), \quad \text{determinant of a block lower triangular matrix}
\end{aligned}$$

so the eigenvalues of the block matrix are those of $\boldsymbol{M}$ together with those of $\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]$, and its spectral radius is $\max\{\rho(\boldsymbol{M}), \rho(\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t])\}$.

Step 4. Bounded affine moments under the strict condition. Suppose $\rho(\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]) < 1$. By the mean-radius inequality in the second-moment characterization (Lemma G.14),

$$\rho(\boldsymbol{M})^2 \le \rho(\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]) < 1.$$

Let $\mathbf{B}$ be the block matrix in (34), let $\mathbf{c} \coloneqq (\mathbb{E}[\boldsymbol{q}_t], \boldsymbol{b}_{\text{quad}})^\top$ be its constant forcing, and write $\mathbf{u}_t \coloneqq (\boldsymbol{z}_t^{(1)}, \boldsymbol{z}_t^{(2)})^\top$. Step 3 gives $\rho(\mathbf{B}) < 1$. Iterating the moment recursion yields

$$\begin{aligned} \mathbf{u}_t &= \mathbf{B}^t \mathbf{u}_0 + \sum_{j=0}^{t-1} \mathbf{B}^j \mathbf{c} && \text{iterate } \mathbf{u}_{t+1} = \mathbf{B}\mathbf{u}_t + \mathbf{c} \\ &\longrightarrow (\boldsymbol{I} - \mathbf{B})^{-1} \mathbf{c}. && \mathbf{B}^t \to 0 \text{ and the Neumann series, } \rho(\mathbf{B}) < 1 \end{aligned}$$

The initial vector $\mathbf{u}_0$ is finite by the second-moment assumption. The trace of the second-moment matrix is $\mathbb{E}\|\boldsymbol{z}_t\|^2$, so convergence of its entries implies convergence, and hence boundedness, of this trace. This proves the affine stability assertion for every initial state of finite second moment independent of the batch sequence. The cross block $\boldsymbol{S}$ already includes any correlation between the current transition and forcing. This completes the proof of Proposition D.3. □

E.3 Proof of Corollary D.4

Proof of Corollary D.4. We center the SGD recurrence at the population stationary point, which removes its mean forcing.

Step 1. The displacement block closes. Take $\beta_1 = 0$, $\boldsymbol{P} = \boldsymbol{I}$ and $\lambda_{\text{WD}} = 0$. Write $\boldsymbol{x}_\star \coloneqq \boldsymbol{w}_\star - \boldsymbol{w}_{t_0}$ and $\bar{\boldsymbol{\epsilon}}_t \coloneqq \boldsymbol{\epsilon}_t - \boldsymbol{x}_\star = \boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star$. The gradient of the surrogate is $\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{\epsilon}_t$, so centering its SGD update gives

$$\begin{aligned} \bar{\boldsymbol{\epsilon}}_{t+1} &= \bar{\boldsymbol{\epsilon}}_t - \eta\big(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}(\bar{\boldsymbol{\epsilon}}_t + \boldsymbol{x}_\star)\big) && \boldsymbol{\epsilon}_t = \bar{\boldsymbol{\epsilon}}_t + \boldsymbol{x}_\star \\ &= (\boldsymbol{I} - \eta\boldsymbol{H}_{\mathcal{B}_t})\bar{\boldsymbol{\epsilon}}_t - \eta(\boldsymbol{g}_{\mathcal{B}_t} + \boldsymbol{H}_{\mathcal{B}_t}\boldsymbol{x}_\star) && \text{collect the centered displacement} \\ &= \boldsymbol{M}_t\bar{\boldsymbol{\epsilon}}_t - \eta\boldsymbol{r}_t. && \text{definitions of } \boldsymbol{M}_t \text{ and } \boldsymbol{r}_t \end{aligned}$$

So the centered displacement is itself a stochastic linear system with transition $\boldsymbol{M}_t$ and forcing $-\eta\boldsymbol{r}_t$. Under Assumption D.1 the pairs $(\boldsymbol{M}_t, -\eta\boldsymbol{r}_t)$ are i.i.d. functions of $(\boldsymbol{H}_{\mathcal{B}_t}, \boldsymbol{g}_{\mathcal{B}_t})$, have finite second moments, and are independent of $\bar{\boldsymbol{\epsilon}}_0 = -\boldsymbol{x}_\star$.

Step 2. Apply Proposition D.3. Apply the proposition to $\bar{\boldsymbol{\epsilon}}_{t+1} = \boldsymbol{M}_t\bar{\boldsymbol{\epsilon}}_t + (-\eta\boldsymbol{r}_t)$. Its mean forcing vanishes,

$$\begin{aligned} \mathbb{E}[-\eta\boldsymbol{r}_t] &= -\eta\big(\mathbb{E}[\boldsymbol{g}_{\mathcal{B}_t}] + \mathbb{E}[\boldsymbol{H}_{\mathcal{B}_t}]\boldsymbol{x}_\star\big) && x_\star \text{ is deterministic} \\ &= -\eta\big(\nabla\mathcal{L}(\boldsymbol{w}_{t_0}) + \nabla^2\mathcal{L}(\boldsymbol{w}_{t_0})(\boldsymbol{w}_\star - \boldsymbol{w}_{t_0})\big) && \text{unbiased mini-batch gradient and Hessian} \\ &= \mathbf{0}, && \text{by (35)} \end{aligned}$$

and its two constants are those of the corollary,

$$\begin{aligned} \mathbb{E}[(-\eta\boldsymbol{r}_t) \otimes \boldsymbol{M}_t + \boldsymbol{M}_t \otimes (-\eta\boldsymbol{r}_t)] &= -\eta\,\mathbb{E}[\boldsymbol{r}_t \otimes \boldsymbol{M}_t + \boldsymbol{M}_t \otimes \boldsymbol{r}_t] = \boldsymbol{S}, && \otimes \text{ is bilinear} \\ \mathsf{vec}\big(\mathbb{E}[(\eta\boldsymbol{r}_t)(\eta\boldsymbol{r}_t)^\top]\big) &= \eta^2\mathsf{vec}\big(\mathbb{E}[\boldsymbol{r}_t\boldsymbol{r}_t^\top]\big) = \boldsymbol{b}_{\text{quad}}. && \text{definition of } \boldsymbol{b}_{\text{quad}} \end{aligned}$$

The two mean identities used above are part of Assumption D.1. The moments $\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t]$ and $\mathsf{vec}(\mathbb{E}[\bar{\boldsymbol{\epsilon}}_t\bar{\boldsymbol{\epsilon}}_t^\top])$ are the $\boldsymbol{\epsilon}_t^{(1)}$ and $\boldsymbol{\epsilon}_t^{(2)}$ of the corollary. Thus (34) is (36), with zero forcing in the mean block. For the generic quadratic of Section 3, apply the same proposition with $\boldsymbol{M}_t = \boldsymbol{I} - \eta\boldsymbol{A}_t$ and $\boldsymbol{q}_t = -\eta(\boldsymbol{A}_t\boldsymbol{x}_\star + \boldsymbol{b}_t)$. Its assumptions follow from the stated iid finite-moment model, and $\mathbb{E}[\boldsymbol{q}_t] = -\eta(\boldsymbol{A}\boldsymbol{x}_\star + \boldsymbol{b}) = 0$. This proves the moment equations of Proposition 3.2. If its second-moment spectral radius is less than one, Step 4 of the generic proof gives finite limiting moments for every deterministic initialization, completing that proposition's affine boundedness claim. This completes the proof of Corollary D.4. □

E.4 Proofs of Lemma D.5 and Corollary D.6

Proof of Lemma D.5. The idea is that the mean AGM dynamics decouple along the eigenvectors of $\boldsymbol{D}$ into 2×2 systems, and that the one along a negative effective curvature has a real eigenvalue larger than one. Let $\mu_1, \ldots, \mu_d$ be the eigenvalues of the population preconditioned Hessian $\boldsymbol{D}$. The corresponding eigenvalues of the effective curvature $\boldsymbol{A} = \boldsymbol{D} + \lambda_{\text{WD}}\boldsymbol{I}$ are $\bar{\mu}_i \coloneqq \mu_i + \lambda_{\text{WD}}$. Since (32) is affine in $\boldsymbol{D}_{\mathcal{B}_t}$, the mean transition $\mathbb{E}[\boldsymbol{M}_t(\eta)]$ is obtained by replacing $\boldsymbol{D}_{\mathcal{B}_t}$ with its mean $\boldsymbol{D}$, so its blocks are $\beta_1\boldsymbol{I}$, $(1-\beta_1)\eta\boldsymbol{D}$, $-\beta_1\boldsymbol{I}$ and $(1-\eta\lambda_{\text{WD}})\boldsymbol{I} - (1-\beta_1)\eta\boldsymbol{D}$. Every block is a polynomial

in $\boldsymbol{D}$, so this transition is block diagonal in an eigenbasis of $\boldsymbol{D}$. Along an eigendirection with eigenvalue μ_i, the mean AGM dynamics reduce to a two-dimensional linear system with transition matrix $\boldsymbol{M}_{\mu_i}(\eta)$, the 2×2 block of $\mathbb{E}[\boldsymbol{M}_t(\eta)]$ along that eigendirection, obtained by replacing $\boldsymbol{D}$ with μ_i in each block:

$$\boldsymbol{M}_{\mu_i}(\eta) = \begin{bmatrix} \beta_1 & (1-\beta_1)\eta\mu_i \\ -\beta_1 & 1 - \eta\lambda_{\mathrm{WD}} - (1-\beta_1)\eta\mu_i \end{bmatrix}.$$

Its characteristic polynomial evaluated at 1 is

$$\begin{aligned} \det(\boldsymbol{I} - \boldsymbol{M}_{\mu_i}(\eta)) &= \det \begin{bmatrix} 1-\beta_1 & -(1-\beta_1)\eta\mu_i \\ \beta_1 & \eta\lambda_{\mathrm{WD}} + (1-\beta_1)\eta\mu_i \end{bmatrix} \\ &= (1-\beta_1)\big(\eta\lambda_{\mathrm{WD}} + (1-\beta_1)\eta\mu_i\big) + \beta_1(1-\beta_1)\eta\mu_i && 2 \times 2 \text{ determinant} \\ &= (1-\beta_1)\eta\lambda_{\mathrm{WD}} + (1-\beta_1)\eta\mu_i\big((1-\beta_1) + \beta_1\big) && \text{collect the } \mu_i \text{ terms} \\ &= (1-\beta_1)\eta\bar{\mu}_i. && \bar{\mu}_i = \mu_i + \lambda_{\mathrm{WD}} \end{aligned}$$

If $\bar{\mu}_i < 0$, this quantity is negative for every $\eta > 0$. The characteristic polynomial $\lambda \mapsto \det(\lambda \boldsymbol{I} - \boldsymbol{M}_{\mu_i}(\eta))$ is monic of degree two, so it is positive for all large λ. Since it is negative at $\lambda = 1$, the intermediate value theorem gives a real root larger than one, that is, an eigenvalue of $\boldsymbol{M}_{\mu_i}(\eta)$ larger than one. The homogeneous mean dynamics therefore have an expanding mode. Choose $\boldsymbol{z}_0$ to be a nonzero right eigenvector of the mean transition for this real eigenvalue $r > 1$. Independence across steps gives $\mathbb{E}[\boldsymbol{z}_t] = r^t \boldsymbol{z}_0$, and Jensen's inequality yields

$$\mathbb{E}\|\boldsymbol{z}_t\|^2 \geq \|\mathbb{E}[\boldsymbol{z}_t]\|^2 = r^{2t}\|\boldsymbol{z}_0\|^2 \longrightarrow \infty.$$

Thus the homogeneous system is not mean-square stable. This completes the proof of Lemma D.5. □

Proof of Corollary D.6. The idea is that the mean recursion is the full-batch recursion, along whose negative eigendirection the mean grows geometrically, and that the second moment dominates the squared mean. Recall from Corollary D.4 that $\boldsymbol{\epsilon}_t^{(1)} = \mathbb{E}[\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star]$ is the mean error and $\boldsymbol{M}_t = \boldsymbol{I} - \eta \boldsymbol{H}_{\mathcal{B}_t}$ is the mini-batch transition matrix, so that the mean recursion in (36) reads $\boldsymbol{\epsilon}_{t+1}^{(1)} = \mathbb{E}[\boldsymbol{M}_t]\boldsymbol{\epsilon}_t^{(1)}$. Recall also that $\boldsymbol{H}_{t_0} = \nabla^2 \mathcal{L}(\boldsymbol{w}_{t_0})$ is the population Hessian at the checkpoint (Appendix D.2), so that $\mathbb{E}[\boldsymbol{M}_t] = \boldsymbol{I} - \eta \boldsymbol{H}_{t_0}$ for every t. If $\boldsymbol{H}_{t_0}\boldsymbol{u} = \lambda \boldsymbol{u}$ with $\lambda < 0$ and $\|\boldsymbol{u}\| = 1$, then

$$\begin{aligned} \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star] &= \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{M}_t]^t \, \mathbb{E}[\boldsymbol{w}_{t_0} - \boldsymbol{w}_\star] && \text{iterate the mean recursion for } t \text{ steps} \\ &= \boldsymbol{u}^\top (\boldsymbol{I} - \eta \boldsymbol{H}_{t_0})^t \, \mathbb{E}[\boldsymbol{w}_{t_0} - \boldsymbol{w}_\star] && \mathbb{E}[\boldsymbol{M}_t] = \boldsymbol{I} - \eta \boldsymbol{H}_{t_0} \\ &= (1-\eta\lambda)^t \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{w}_{t_0} - \boldsymbol{w}_\star]. && \boldsymbol{u}^\top \boldsymbol{H}_{t_0} = \lambda \boldsymbol{u}^\top, \text{ as } \boldsymbol{H}_{t_0} \text{ is symmetric} \end{aligned}$$

Because $1 - \eta\lambda > 1$, a nonzero initial component along $\boldsymbol{u}$ grows geometrically. The second moment then diverges for this initialization:

$$\begin{aligned} \mathbb{E}\|\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star\|^2 &\geq \|\mathbb{E}[\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star]\|^2 && \text{by Jensen's inequality} \\ &\geq \big(\boldsymbol{u}^\top \mathbb{E}[\boldsymbol{w}_{t_0+t} - \boldsymbol{w}_\star]\big)^2 && \text{by the Cauchy–Schwarz inequality} \\ &= (1-\eta\lambda)^{2t}\big(\boldsymbol{u}^\top \mathbb{E}[\boldsymbol{w}_{t_0} - \boldsymbol{w}_\star]\big)^2 && \text{by the display above} \\ &\to \infty. && 1 - \eta\lambda > 1 \end{aligned}$$

The same computation with $\boldsymbol{r}_t \equiv \boldsymbol{0}$ shows that the homogeneous system $\bar{\boldsymbol{\epsilon}}_{t+1} = \boldsymbol{M}_t \bar{\boldsymbol{\epsilon}}_t$ is not mean-square stable. The population quadratic evaluated at $\boldsymbol{w}_\star + a\boldsymbol{u}$ is a quadratic polynomial in a whose a^2 coefficient is $\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{H}_{t_0}\boldsymbol{u} = \lambda/2 < 0$, so it is unbounded below. This completes the proof of Corollary D.6. □

E.5 Proof of Corollary D.8

Proof of Corollary D.8. We use a PSD eigenmatrix of the adjoint second-moment map to identify an initialization with geometric second-moment growth. Recall from Appendix D.2 that $\boldsymbol{M}_t^+(\eta)$ is the transition of the projected homogeneous system, acting on the retained momentum and

displacement coordinates, and that $\boldsymbol{K}^{+}(\eta) = \mathbb{E}[\boldsymbol{M}_t^{+}(\eta) \otimes \boldsymbol{M}_t^{+}(\eta)]$. By the unique stability threshold (Theorem D.7), $\rho(\boldsymbol{K}^{+}(\eta)) > 1$ for every $\eta > \eta_\star$. Write this radius as $r > 1$ and define the second-moment map on symmetric matrices by

$$\mathcal{T}(\boldsymbol{X}) \coloneqq \mathbb{E}[\boldsymbol{M}_t^{+}(\eta) \boldsymbol{X} \boldsymbol{M}_t^{+}(\eta)^\top].$$

Restriction to symmetric matrices preserves the spectral radius, so $\rho(\mathcal{T}) = r$ (see the discussion after Lemma G.14 in Appendix G.6). The adjoint map $\mathcal{T}^*(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{M}_t^{+}(\eta)^\top \boldsymbol{X}\, \boldsymbol{M}_t^{+}(\eta)]$ preserves PSD matrices and has the same spectral radius as $\mathcal{T}$, so by the Perron–Frobenius theorem for cone-preserving maps (Lemma G.16) it has a nonzero PSD eigenmatrix $\boldsymbol{W}$ with $\mathcal{T}^*(\boldsymbol{W}) = r\boldsymbol{W}$. Choose a deterministic initial state $\boldsymbol{z}_0$ of the projected system such that $\boldsymbol{z}_0^\top \boldsymbol{W} \boldsymbol{z}_0 > 0$, for instance an eigenvector of $\boldsymbol{W}$ with positive eigenvalue. Its second moment $\boldsymbol{X}_t = \mathbb{E}[\boldsymbol{z}_t \boldsymbol{z}_t^\top]$ satisfies $\boldsymbol{X}_t = \mathcal{T}^t(\boldsymbol{z}_0 \boldsymbol{z}_0^\top)$, by the recursion $\mathbb{E}[\boldsymbol{z}_{t+1} \boldsymbol{z}_{t+1}^\top] = \mathcal{T}(\mathbb{E}[\boldsymbol{z}_t \boldsymbol{z}_t^\top])$ of the second-moment characterization (Lemma G.14). Hence

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{z}_t\|^2 &= \operatorname{tr}(\boldsymbol{X}_t) && \|\boldsymbol{z}_t\|^2 = \operatorname{tr}(\boldsymbol{z}_t \boldsymbol{z}_t^\top);\ \mathbb{E} \text{ and } \operatorname{tr} \text{ commute} \\
&\geq \frac{\operatorname{tr}(\boldsymbol{W} \boldsymbol{X}_t)}{\|\boldsymbol{W}\|} && \boldsymbol{X}_t \succeq 0 \text{ and } \boldsymbol{0} \preceq \boldsymbol{W} \preceq \|\boldsymbol{W}\| \boldsymbol{I} \\
&= \frac{\operatorname{tr}\big(\boldsymbol{W}\, \mathcal{T}^t(\boldsymbol{z}_0 \boldsymbol{z}_0^\top)\big)}{\|\boldsymbol{W}\|} && \boldsymbol{X}_t = \mathcal{T}^t(\boldsymbol{z}_0 \boldsymbol{z}_0^\top) \\
&= \frac{\operatorname{tr}\big((\mathcal{T}^*)^t(\boldsymbol{W})\, \boldsymbol{z}_0 \boldsymbol{z}_0^\top\big)}{\|\boldsymbol{W}\|} && \text{definition of the adjoint, applied } t \text{ times} \\
&= \frac{r^t \operatorname{tr}(\boldsymbol{W} \boldsymbol{z}_0 \boldsymbol{z}_0^\top)}{\|\boldsymbol{W}\|} && \mathcal{T}^*(\boldsymbol{W}) = r\boldsymbol{W} \\
&= \frac{r^t \boldsymbol{z}_0^\top \boldsymbol{W} \boldsymbol{z}_0}{\|\boldsymbol{W}\|} \longrightarrow \infty. && \text{cyclicity of the trace; } r > 1 \text{ and } \boldsymbol{z}_0^\top \boldsymbol{W} \boldsymbol{z}_0 > 0
\end{aligned}$$

For SGD, $\beta_1 = 0$, $\boldsymbol{P} = \boldsymbol{I}$ and $\lambda_{\mathrm{WD}} = 0$, the projected state reduces to the restricted displacement $\boldsymbol{x}_t \in \operatorname{range}(\Pi_{t_0}^{>0})$ with transition $\boldsymbol{M}_t^{+}(\eta) = \boldsymbol{I} - \eta \Pi_{t_0}^{>0} \boldsymbol{H}_{\mathcal{B}_t} \Pi_{t_0}^{>0}$, and the same argument gives the exploding initial displacement of the SGD case in Appendix D.2. This completes the proof of Corollary D.8. □

E.6 Proof of the AGM Critical LR Characterization

Proof of Proposition B.3. We show that every LR below the spectral threshold has bounded affine moments, while no LR above it does so for every initialization. This identifies the supremum without making an assumption about the endpoint. Recall the projected affine system (16),

$$\boldsymbol{z}_{t+1}^{+} = \boldsymbol{M}_t^{+}(\eta)\boldsymbol{z}_t^{+} + \boldsymbol{q}_t^{+}, \qquad \boldsymbol{K}^{+}(\eta) = \mathbb{E}[\boldsymbol{M}_t^{+}(\eta) \otimes \boldsymbol{M}_t^{+}(\eta)].$$

The pairs $(\boldsymbol{M}_t^{+}(\eta), \boldsymbol{q}_t^{+})$ are i.i.d. with finite second moments by Assumption D.1. The transition and forcing within a pair may be correlated. The retained effective curvature is positive definite. By Corollary G.4 and the change of state in Appendix G.6, $\rho(\boldsymbol{K}^{+}(\eta))$ is less than, equal to, or greater than one below, at, or above a threshold $\eta_{\mathrm{crit}} \in (0, \infty)$.

Step 1. LRs below the spectral threshold. If $0 < \eta < \eta_{\mathrm{crit}}$, then $\rho(\boldsymbol{K}^{+}(\eta)) < 1$. The affine moment recursion in Proposition D.3 therefore gives

$$\sup_{t \geq 0} \mathbb{E}\|\boldsymbol{z}_t^{+}\|^2 < \infty \qquad \text{for every deterministic } \boldsymbol{z}_0^{+}.$$

Thus every such LR belongs to the stable set in Definition B.2.

Step 2. LRs above the spectral threshold. Let $\eta > \eta_{\mathrm{crit}}$, so that $\rho(\boldsymbol{K}^{+}(\eta)) > 1$. The Perron–Frobenius argument in Appendix E.5 then gives a deterministic initial vector $\boldsymbol{u}$ whose homogeneous second moment diverges. Run two affine trajectories, $\boldsymbol{z}_t^{+}$ and $\widehat{\boldsymbol{z}}_t^{+}$, with the same pairs $(\boldsymbol{M}_t^{+}(\eta), \boldsymbol{q}_t^{+})$ and initial states $\boldsymbol{z}_0^{+} = \boldsymbol{u}$ and $\widehat{\boldsymbol{z}}_0^{+} = 0$. Their difference $\boldsymbol{d}_t := \boldsymbol{z}_t^{+} - \widehat{\boldsymbol{z}}_t^{+}$ satisfies

$$\begin{aligned} \boldsymbol{d}_{t+1} &= \big(\boldsymbol{M}_t^{+}(\eta)\boldsymbol{z}_t^{+} + \boldsymbol{q}_t^{+}\big) - \big(\boldsymbol{M}_t^{+}(\eta)\widehat{\boldsymbol{z}}_t^{+} + \boldsymbol{q}_t^{+}\big) && \text{use the same batches and forcing} \\ &= \boldsymbol{M}_t^{+}(\eta)\boldsymbol{d}_t, \qquad \boldsymbol{d}_0 = \boldsymbol{u}. && \text{the forcing cancels pathwise} \end{aligned}$$

If the affine system were bounded in mean square for every deterministic start, then

$$\begin{aligned} \sup_{t \geq 0} \mathbb{E}\|\boldsymbol{d}_t\|^2 &\leq 2 \sup_{t \geq 0} \mathbb{E}\|\boldsymbol{z}_t^{+}\|^2 + 2 \sup_{t \geq 0} \mathbb{E}\|\widehat{\boldsymbol{z}}_t^{+}\|^2 && \|\boldsymbol{a} - \boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2 \\ &< \infty, && \text{boundedness for the two chosen starts} \end{aligned}$$

contradicting the choice of $\boldsymbol{u}$. Thus no $\eta > \eta_{\mathrm{crit}}$ belongs to the stable set. Together, the two steps give

$$(0, \eta_{\mathrm{crit}}) \subseteq \{\eta > 0 : (16) \text{ is mean square stable}\} \subseteq (0, \eta_{\mathrm{crit}}].$$

The middle set has supremum $\eta_\star$ by Definition B.2, so $\eta_\star = \eta_{\mathrm{crit}}$. The spectral characterization and the three cases in (20) now follow from Corollary G.4. At the endpoint, forcing can change boundedness without changing this supremum. For example, $x_{t+1} = -x_t$ has bounded squared iterates, whereas $x_{t+1} = -x_t + \xi_t$ with independent uniform signs ξ_t satisfies

$$\begin{aligned} \mathbb{E}[x_{t+1}^2] &= \mathbb{E}[x_t^2] - 2\mathbb{E}[x_t\xi_t] + \mathbb{E}[\xi_t^2] \\ &= \mathbb{E}[x_t^2] + 1 = x_0^2 + t + 1. \end{aligned}$$

The second line uses independence, $\mathbb{E}[\xi_t] = 0$ and $\xi_t^2 = 1$, then iterates the recursion. Both have second-moment spectral radius one. This completes the proof of Proposition B.3. □

F Proof of the Power-Iteration Bound

F.1 The Bisection Bound

We prove Theorem D.12 by bounding the two possible query errors, then tracking the bisection bracket. Below the target shift, we show that the bottom-eigenvector component grows past the cutoff with high probability. Above the target, we bound the probability of ever crossing the cutoff.

Recall that $\boldsymbol{A} = \boldsymbol{D}_{t_0} + \lambda_{\mathrm{WD}}\boldsymbol{I}$ is the effective curvature and $\boldsymbol{\Delta}_t = \boldsymbol{A}_{\mathcal{B}_t} - \boldsymbol{A}$ is its centered mini-batch noise. A query at shift λ, with inner stepsize η_{p} and norm cutoff B, follows

$$\boldsymbol{x}_{t+1} = \big[\boldsymbol{M}(\lambda) - \eta_{\mathrm{p}}\boldsymbol{\Delta}_t\big]\boldsymbol{x}_t, \qquad \boldsymbol{M}(\lambda) = \boldsymbol{I} - \eta_{\mathrm{p}}(\boldsymbol{A} + \lambda\boldsymbol{I}), \qquad \|\boldsymbol{x}_0\| = 1.$$

Under Assumption D.10, $\lambda_\star = -\lambda_{\min}(\boldsymbol{A})$. For this proof, define the largest eigenvalue of the mean transition and the first exit time by

$$\rho(\lambda) \coloneqq 1 + \eta_{\mathrm{p}}(\lambda_\star - \lambda),$$
$$\tau \coloneqq \inf\{t \geq 0 : \|\boldsymbol{x}_t\| > B\}, \qquad \inf \varnothing = \infty.$$

A query of horizon T_{p} explodes precisely when $\tau \leq T_{\mathrm{p}}$. The filtration $\mathcal{F}_t$ includes all previous queries and the first t observations of the current query, as defined in Appendix D.3. The query bounds below hold conditionally on the previous queries, because Assumption D.9 gives conditional centering and the same noise bound under this conditioning.

The next two lemmas control the query errors below and above $\lambda_\star$. Their proofs are given in Appendices F.2 and F.3. In the first lemma, write $\rho = \rho(\lambda)$.

Lemma F.1 (Explosion below $\lambda_\star$)**.** *Let Assumptions D.9 and D.10 hold, let $\eta_{\mathrm{p}} > 0$, $B > 1$, $\lambda \in [0, \lambda_\star)$, and let $T_{\mathrm{p}} \geq 1$ be an integer with $B\rho^{-T_{\mathrm{p}}} < \varepsilon$. Then*

$$\mathbb{P}\{\tau > T_{\mathrm{p}}\} \leq \exp\left(-\frac{(\varepsilon - B\rho^{-T_{\mathrm{p}}})^2\,(\rho^2 - 1)}{2\eta_{\mathrm{p}}^2\sigma^2 B^2}\right). \tag{57}$$

In particular, let $g > 0$ and let e_{p} be the finite-horizon term of (43). If $\lambda \leq \lambda_\star - \max\{g, e_{\mathrm{p}}\}$, then

$$\mathbb{P}\{\tau > T_{\mathrm{p}}\} \leq \exp\Big(-\frac{\varepsilon^2 g}{4\eta_{\mathrm{p}}\sigma^2 B^2}\Big), \tag{58}$$

and for $g = g_-$ the right-hand side equals δ/N.

Lemma F.2 (No explosion above $\lambda_\star$)**.** *Let Assumptions D.9 to D.11 hold, let $B > 1$, and fix $\lambda \in [\lambda_\star, \Lambda]$. With $a \coloneqq \eta_{\mathrm{p}}(\lambda - \lambda_\star)$ and $h \coloneqq \eta_{\mathrm{p}}\sigma$, suppose $a > h^2$. Then*

$$\mathbb{P}\big\{\exists\, t \geq 0 : \|\boldsymbol{x}_t\| > B\big\} \leq B^{-(a-h^2)/h^2}.$$

If $a \geq 2h^2$, the right-hand side is at most $B^{-(\lambda-\lambda_\star)/(2\eta_{\mathrm{p}}\sigma^2)}$. In particular, when $\lambda \geq \lambda_\star + g_+$, with g_+ from (43), the probability is at most δ/N.

Proof of Theorem D.12. Recall that g_- and g_+ in (43) are the noise margins below and above $\lambda_\star$, respectively. The finite horizon contributes $e_{\mathrm{p}} = ((2B/\varepsilon)^{1/T_{\mathrm{p}}} - 1)/\eta_{\mathrm{p}}$. Fix an integer $T_{\mathrm{p}} \geq 1$.

Step 1. Control the queries outside the error margins.

By the gap form (58) of Lemma F.1 with $g = g_-$, a query at $\lambda \leq \lambda_\star - \max\{g_-, e_{\mathrm{p}}\}$ explodes with probability at least $1 - \delta/N$. By Lemma F.2, a query at $\lambda \geq \lambda_\star + g_+$ does not explode with probability at least $1 - \delta/N$.

For the kth query, let $\mathcal{G}_{k-1}$ contain all observations of the preceding queries, and let E_k be the event that this query returns the wrong outcome in either of these two regions. Conditionally on $\mathcal{G}_{k-1}$, the queried shift is fixed and Assumption D.9 still holds. The two regions are disjoint, so the query can incur only one of the two errors. Therefore

$$\mathbb{P}(E_k \mid \mathcal{G}_{k-1}) \leq \frac{\delta}{N} \qquad \text{almost surely.}$$

Taking expectations and then a union bound gives

$$\begin{aligned}\mathbb{P}\Big(\bigcup_{k=1}^{N} E_k\Big) &\le \sum_{k=1}^{N} \mathbb{P}(E_k) && \text{union bound}\\ &= \sum_{k=1}^{N} \mathbb{E}\big[\mathbb{P}(E_k \mid \mathcal{G}_{k-1})\big] && \text{tower property}\\ &\le \sum_{k=1}^{N} \frac{\delta}{N} = \delta. && \text{the conditional bound}\end{aligned}$$

Thus, with probability at least $1-\delta$, every query in either region has the stated outcome. Queries strictly between the two regions may return either outcome.

Step 2. The bracket.

On this event, an exploding query has $\lambda < \lambda_\star + g_+$ and a non-exploding query has $\lambda > \lambda_\star - \max\{g_-, e_{\mathrm{p}}\}$. Every lower endpoint of the bracket is either 0 or an exploding query, hence at most $\lambda_\star + g_+$. Every upper endpoint is either Λ, which is at least $\lambda_\star$ by Assumption D.10, or a non-exploding query, hence at least $\lambda_\star - \max\{g_-, e_{\mathrm{p}}\}$. After N halvings the bracket has length $\Lambda/2^N$, so

$$\lambda_\star - \max\{g_-, e_{\mathrm{p}}\} \le \widehat{\lambda} = \lambda_{\mathrm{lo}} + \frac{\Lambda}{2^N} \le \lambda_\star + g_+ + \frac{\Lambda}{2^N},$$

which is (44).

Step 3. The crossover.

By the definition of e_{p},

$$\begin{aligned}e_{\mathrm{p}} \le g_- &\iff (2B/\varepsilon)^{1/T_{\mathrm{p}}} \le 1 + \eta_{\mathrm{p}} g_-\\ &\iff \frac{\log(2B/\varepsilon)}{T_{\mathrm{p}}} \le \log(1+\eta_{\mathrm{p}} g_-)\\ &\iff T_{\mathrm{p}} \ge \frac{\log(2B/\varepsilon)}{\log(1+\eta_{\mathrm{p}} g_-)}.\end{aligned}$$

The second equivalence takes logarithms of positive quantities. Since T_{p} is an integer, this is equivalent to the ceiling condition in (45). This completes the proof of Theorem D.12. □

F.2 Proof of the Explosion Bound

Fix a query at a shift $\lambda \in [0, \lambda_\star)$. Recall that $\rho = \rho(\lambda) = 1 + \eta_{\mathrm{p}}(\lambda_\star - \lambda) > 1$ is the largest eigenvalue of $\boldsymbol{M}(\lambda)$, attained at $\boldsymbol{u}_{\min}$, so that $\boldsymbol{u}_{\min}^\top \boldsymbol{M}(\lambda) = \rho\, \boldsymbol{u}_{\min}^\top$, and that τ is the first time at which $\|\boldsymbol{x}_t\|$ exceeds the cutoff B. Choose the sign of $\boldsymbol{u}_{\min}$ so that $\boldsymbol{u}_{\min}^\top \boldsymbol{x}_0 \ge \varepsilon$, which Assumption D.10 allows.

Proof of Lemma F.1. The idea is to divide the bottom-eigenvector component by its mean growth factor ρ^t and stop it at τ. The resulting martingale has increment bounds whose squares form a summable geometric sequence. The Azuma–Hoeffding inequality then bounds the probability that this component falls as far as the event $\tau > T_{\mathrm{p}}$ would require.

Step 1. A stopped martingale with summable increment bounds.

Set $Z_t := \rho^{-t}\, \boldsymbol{u}_{\min}^\top \boldsymbol{x}_t$. By the iteration (40),

$$\begin{aligned}Z_{t+1} - Z_t &= \rho^{-(t+1)}\, \boldsymbol{u}_{\min}^\top \big(\boldsymbol{M}(\lambda) - \eta_{\mathrm{p}} \boldsymbol{\Delta}_t\big)\boldsymbol{x}_t - \rho^{-t}\, \boldsymbol{u}_{\min}^\top \boldsymbol{x}_t\\ &= \big(\rho^{-(t+1)}\rho - \rho^{-t}\big)\boldsymbol{u}_{\min}^\top \boldsymbol{x}_t - \eta_{\mathrm{p}} \rho^{-(t+1)} \boldsymbol{u}_{\min}^\top \boldsymbol{\Delta}_t \boldsymbol{x}_t\\ &= -\eta_{\mathrm{p}} \rho^{-(t+1)}\, \boldsymbol{u}_{\min}^\top \boldsymbol{\Delta}_t \boldsymbol{x}_t.\end{aligned}$$

The second equality uses $\boldsymbol{u}_{\min}^\top \boldsymbol{M}(\lambda) = \rho \boldsymbol{u}_{\min}^\top$. Since $\boldsymbol{x}_t$ is $\mathcal{F}_t$-measurable, the conditional increment is

$$\begin{aligned}\mathbb{E}[Z_{t+1} - Z_t \mid \mathcal{F}_t] &= -\eta_{\mathrm{p}} \rho^{-(t+1)} \boldsymbol{u}_{\min}^\top \mathbb{E}[\boldsymbol{\Delta}_t \mid \mathcal{F}_t] \boldsymbol{x}_t && \text{conditional measurability} \\ &= 0. && \text{Assumption D.9}\end{aligned}$$

The bounded noise makes the state, and hence Z_t, integrable at each finite time. Thus $(Z_t)_{t \geq 0}$ is a martingale with $Z_0 = \boldsymbol{u}_{\min}^\top \boldsymbol{x}_0 \geq \varepsilon$.

We stop it at τ. The event $\{t < \tau\} = \{\max_{s \leq t} \|\boldsymbol{x}_s\| \leq B\}$ is $\mathcal{F}_t$-measurable. The stopped increments therefore satisfy

$$\begin{aligned}Z_{(t+1)\wedge\tau} - Z_{t\wedge\tau} &= \mathbf{1}_{\{t<\tau\}} (Z_{t+1} - Z_t), \\ \mathbb{E}[Z_{(t+1)\wedge\tau} - Z_{t\wedge\tau} \mid \mathcal{F}_t] &= \mathbf{1}_{\{t<\tau\}} \mathbb{E}[Z_{t+1} - Z_t \mid \mathcal{F}_t] \\ &= 0.\end{aligned}$$

Thus the stopped process is again a martingale. On $\{t < \tau\}$ we have $\|\boldsymbol{x}_t\| \leq B$, so

$$\begin{aligned}\left|Z_{(t+1)\wedge\tau} - Z_{t\wedge\tau}\right| &= \mathbf{1}_{\{t<\tau\}} \eta_{\mathrm{p}} \rho^{-(t+1)} |\boldsymbol{u}_{\min}^\top \boldsymbol{\Delta}_t \boldsymbol{x}_t| \\ &\leq \mathbf{1}_{\{t<\tau\}} \eta_{\mathrm{p}} \rho^{-(t+1)} \|\boldsymbol{u}_{\min}\| \, \|\boldsymbol{\Delta}_t\| \, \|\boldsymbol{x}_t\| && \text{Cauchy–Schwarz} \\ &\leq \eta_{\mathrm{p}} \sigma B \, \rho^{-(t+1)}. && \|\boldsymbol{x}_t\| \leq B \text{ before } \tau\end{aligned}$$

Since $\rho > 1$, extending the series of positive terms and summing it gives

$$\begin{aligned}\sum_{t=0}^{T_{\mathrm{p}}-1} \left(\eta_{\mathrm{p}} \sigma B \rho^{-(t+1)}\right)^2 &= \eta_{\mathrm{p}}^2 \sigma^2 B^2 \sum_{t=1}^{T_{\mathrm{p}}} \rho^{-2t} && \text{reindex} \\ &\leq \eta_{\mathrm{p}}^2 \sigma^2 B^2 \sum_{t=1}^{\infty} \rho^{-2t} && \text{nonnegative summands} \\ &= \eta_{\mathrm{p}}^2 \sigma^2 B^2 \cdot \frac{\rho^{-2}}{1 - \rho^{-2}} && \text{geometric series, } \rho > 1 \\ &= \frac{\eta_{\mathrm{p}}^2 \sigma^2 B^2}{\rho^2 - 1}. && \text{multiply by } \rho^2/\rho^2\end{aligned}$$

Step 2. The Azuma–Hoeffding bound.

Suppose $\tau > T_{\mathrm{p}}$. Then $T_{\mathrm{p}} \wedge \tau = T_{\mathrm{p}}$ and $\|\boldsymbol{x}_{T_{\mathrm{p}}}\| \leq B$, so by the Cauchy–Schwarz inequality

$$\begin{aligned}Z_{T_{\mathrm{p}}\wedge\tau} = Z_{T_{\mathrm{p}}} &= \rho^{-T_{\mathrm{p}}} \boldsymbol{u}_{\min}^\top \boldsymbol{x}_{T_{\mathrm{p}}} \\ &\leq \rho^{-T_{\mathrm{p}}} \|\boldsymbol{x}_{T_{\mathrm{p}}}\| \\ &\leq B \rho^{-T_{\mathrm{p}}}.\end{aligned}$$

Since $Z_0 \geq \varepsilon$, it follows that

$$Z_{T_{\mathrm{p}}\wedge\tau} - Z_0 \leq B\rho^{-T_{\mathrm{p}}} - \varepsilon < 0.$$

By the Azuma–Hoeffding inequality (Lemma F.5), applied to the martingale $(Z_{t\wedge\tau})$ with the increment bounds above and with $r = \varepsilon - B\rho^{-T_{\mathrm{p}}} > 0$,

$$\begin{aligned}\mathbb{P}\{\tau > T_{\mathrm{p}}\} &\leq \mathbb{P}\left\{Z_{T_{\mathrm{p}}\wedge\tau} - Z_0 \leq -(\varepsilon - B\rho^{-T_{\mathrm{p}}})\right\} \\ &\leq \exp\left(-\frac{(\varepsilon - B\rho^{-T_{\mathrm{p}}})^2}{2\eta_{\mathrm{p}}^2 \sigma^2 B^2/(\rho^2 - 1)}\right),\end{aligned}$$

which is (57).

Step 3. The gap form.

Suppose $\lambda \le \lambda_\star - \max\{g, e_{\mathrm{p}}\}$. Recall that $e_{\mathrm{p}} = ((2B/\varepsilon)^{1/T_{\mathrm{p}}} - 1)/\eta_{\mathrm{p}}$, so $1 + \eta_{\mathrm{p}} e_{\mathrm{p}} = (2B/\varepsilon)^{1/T_{\mathrm{p}}}$. Then

$$\begin{aligned} \rho &= 1 + \eta_{\mathrm{p}}(\lambda_\star - \lambda) \ge 1 + \eta_{\mathrm{p}} \max\{g, e_{\mathrm{p}}\}, \\ B\rho^{-T_{\mathrm{p}}} &\le B\,(1 + \eta_{\mathrm{p}} e_{\mathrm{p}})^{-T_{\mathrm{p}}} \\ &= B \cdot \frac{\varepsilon}{2B} = \frac{\varepsilon}{2}, \\ \rho^2 - 1 &= (\rho - 1)(\rho + 1) \ge 2(\rho - 1) \ge 2\eta_{\mathrm{p}} g. \end{aligned}$$

The last line uses $\rho \ge 1$ and $\rho \ge 1 + \eta_{\mathrm{p}} g$. Substituting these two bounds into the exponent of (57),

$$\begin{aligned} \frac{(\varepsilon - B\rho^{-T_{\mathrm{p}}})^2(\rho^2 - 1)}{2\eta_{\mathrm{p}}^2 \sigma^2 B^2} &\ge \frac{(\varepsilon/2)^2 \cdot 2\eta_{\mathrm{p}} g}{2\eta_{\mathrm{p}}^2 \sigma^2 B^2} \\ &= \frac{\varepsilon^2 g}{4\eta_{\mathrm{p}} \sigma^2 B^2}, \end{aligned}$$

which gives (58). Recall $g_- = 4\eta_{\mathrm{p}} \sigma^2 B^2 \log(N/\delta)/\varepsilon^2$. Substituting this value gives

$$\begin{aligned} \frac{\varepsilon^2 g_-}{4\eta_{\mathrm{p}} \sigma^2 B^2} &= \frac{\varepsilon^2}{4\eta_{\mathrm{p}} \sigma^2 B^2} \frac{4\eta_{\mathrm{p}} \sigma^2 B^2}{\varepsilon^2} \log \frac{N}{\delta} \\ &= \log \frac{N}{\delta}, \\ \mathbb{P}\{\tau > T_{\mathrm{p}}\} &\le \exp\left(-\log \frac{N}{\delta}\right) = \frac{\delta}{N}. \end{aligned}$$

This completes the proof of Lemma F.1. □

The horizon enters (57) through $B\rho^{-T_{\mathrm{p}}}$. The definition of e_{p} ensures that this quantity is at most $\varepsilon/2$.

F.3 Proof of the No-Explosion Bound

Fix a query at $\lambda \in [\lambda_\star, \Lambda]$. In the adaptive application, we condition on $\mathcal{F}_0$, which contains the previous queries, so this shift is fixed throughout the proof. Recall that the update is $\boldsymbol{x}_{t+1} = [\boldsymbol{M}(\lambda) - \eta_{\mathrm{p}} \boldsymbol{\Delta}_t]\boldsymbol{x}_t$, where $\boldsymbol{M}(\lambda) = \boldsymbol{I} - \eta_{\mathrm{p}}(\boldsymbol{A} + \lambda \boldsymbol{I})$ and $\boldsymbol{\Delta}_t = \boldsymbol{A}_{\mathcal{B}_t} - \boldsymbol{A}$ has conditional mean zero and norm at most σ. Define the contraction margin and the noise bound per step by

$$a \coloneqq \eta_{\mathrm{p}}(\lambda - \lambda_\star), \qquad h \coloneqq \eta_{\mathrm{p}} \sigma.$$

The margin lies in $[0, 1]$. Indeed, since $\lambda_\star = -\lambda_{\min}(\boldsymbol{A})$,

$$\begin{aligned} 0 \le a &= \eta_{\mathrm{p}}\big(\lambda + \lambda_{\min}(\boldsymbol{A})\big) \\ &\le \eta_{\mathrm{p}}\big(\Lambda + \lambda_{\max}(\boldsymbol{A})\big) \\ &\le 1. \qquad \text{Assumption D.11} \end{aligned}$$

The same assumption bounds the eigenvalues of the mean transition:

$$0 \le 1 - \eta_{\mathrm{p}}\big(\lambda_{\max}(\boldsymbol{A}) + \Lambda\big) \le \lambda_i\big(\boldsymbol{M}(\lambda)\big) \le 1 - \eta_{\mathrm{p}}\big(\lambda_{\min}(\boldsymbol{A}) + \lambda\big) = 1 - a.$$

The upper endpoint is attained at $\boldsymbol{u}_{\min}$, so

$$\|\boldsymbol{M}(\lambda)\| = 1 - a. \tag{59}$$

We use this contraction to control the chance that curvature noise pushes the trajectory past the cutoff.

Proof of Lemma F.2. We choose a positive power of the norm whose conditional expectation cannot increase. Hoeffding's lemma controls the curvature noise in one step, and Ville's maximal inequality then controls the probability of ever crossing the cutoff.

Step 1. Expand the one-step change in the norm.

For $\boldsymbol{x}_t \neq \mathbf{0}$, set $\hat{\boldsymbol{x}}_t \coloneqq \boldsymbol{x}_t/\|\boldsymbol{x}_t\|$. Define

$$\xi_t \coloneqq -2\eta_{\mathrm{p}}\langle \boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t, \boldsymbol{\Delta}_t\hat{\boldsymbol{x}}_t\rangle.$$

If $\boldsymbol{x}_t = \mathbf{0}$, set both $\hat{\boldsymbol{x}}_t$ and ξ_t to zero. Then $\boldsymbol{x}_{t+1} = \mathbf{0}$ and the norm bounds below remain valid. For $\boldsymbol{x}_t \neq \mathbf{0}$, expanding the update gives

$$\begin{aligned}
\frac{\|\boldsymbol{x}_{t+1}\|^2}{\|\boldsymbol{x}_t\|^2} &= \|\boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t - \eta_{\mathrm{p}}\boldsymbol{\Delta}_t\hat{\boldsymbol{x}}_t\|^2 && \text{the update} \\
&= \|\boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t\|^2 - 2\eta_{\mathrm{p}}\langle \boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t, \boldsymbol{\Delta}_t\hat{\boldsymbol{x}}_t\rangle \\
&\quad + \eta_{\mathrm{p}}^2\|\boldsymbol{\Delta}_t\hat{\boldsymbol{x}}_t\|^2 && \text{expand the square} \\
&= \|\boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t\|^2 + \xi_t + \eta_{\mathrm{p}}^2\|\boldsymbol{\Delta}_t\hat{\boldsymbol{x}}_t\|^2 && \text{the definition of } \xi_t \\
&\leq (1-a)^2 + \xi_t + h^2 && (59) \text{ and } \|\boldsymbol{\Delta}_t\| \leq \sigma \\
&= 1 - 2a + a^2 + \xi_t + h^2 && \text{expand } (1-a)^2 \\
&\leq 1 + \xi_t - (a - h^2) && 0 \leq a \leq 1 \\
&\leq \exp\bigl(\xi_t - (a - h^2)\bigr) && 1 + y \leq e^y
\end{aligned}$$

The conditional mean of ξ_t is zero. The queried shift and $\hat{\boldsymbol{x}}_t$ are $\mathcal{F}_t$-measurable, so Assumption D.9 gives

$$\begin{aligned}
\mathbb{E}[\xi_t \mid \mathcal{F}_t] &= -2\eta_{\mathrm{p}}\mathbb{E}\bigl[\langle \boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t, \boldsymbol{\Delta}_t\hat{\boldsymbol{x}}_t\rangle \mid \mathcal{F}_t\bigr] \\
&= -2\eta_{\mathrm{p}}\,\langle \boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t, \mathbb{E}[\boldsymbol{\Delta}_t \mid \mathcal{F}_t]\hat{\boldsymbol{x}}_t\rangle && \text{conditional measurability} \\
&= 0. && \text{Assumption D.9}
\end{aligned}$$

Its magnitude is bounded by $2h$:

$$\begin{aligned}
|\xi_t| &\leq 2\eta_{\mathrm{p}}\|\boldsymbol{M}(\lambda)\hat{\boldsymbol{x}}_t\|\|\boldsymbol{\Delta}_t\hat{\boldsymbol{x}}_t\| && \text{Cauchy–Schwarz} \\
&\leq 2\eta_{\mathrm{p}}(1-a)\sigma && \|\hat{\boldsymbol{x}}_t\| \leq 1 \\
&\leq 2\eta_{\mathrm{p}}\sigma = 2h. && 0 \leq a \leq 1
\end{aligned}$$

Step 2. Construct a nonnegative supermartingale.

Since $a > h^2$ and $h = \eta_{\mathrm{p}}\sigma > 0$, define

$$p \coloneqq \frac{a - h^2}{2h^2} > 0, \qquad Y_t \coloneqq \|\boldsymbol{x}_t\|^{2p}.$$

The preceding norm-ratio bound implies

$$Y_{t+1} \leq Y_t \exp\bigl(p\xi_t - p(a - h^2)\bigr).$$

When $\boldsymbol{x}_t = \mathbf{0}$, the next state is also zero, so this inequality holds with both sides zero without dividing by $\|\boldsymbol{x}_t\|$. By Hoeffding's lemma (Lemma F.3), applied conditionally to ξ_t with $|\xi_t| \leq 2h$,

$$\begin{aligned}
\mathbb{E}[Y_{t+1} \mid \mathcal{F}_t] &\leq \mathbb{E}\bigl[Y_t e^{p\xi_t - p(a-h^2)} \mid \mathcal{F}_t\bigr] && \text{the one-step bound} \\
&= Y_t e^{-p(a-h^2)}\mathbb{E}[e^{p\xi_t} \mid \mathcal{F}_t] && Y_t \text{ is } \mathcal{F}_t\text{-measurable} \\
&\leq Y_t \exp\left(-p(a-h^2) + \frac{p^2(2h)^2}{2}\right) && \text{Hoeffding's lemma} \\
&= Y_t \exp\left(-\frac{(a-h^2)^2}{2h^2} + \frac{(a-h^2)^2}{2h^2}\right) && \text{the definition of } p \\
&= Y_t. && (60)
\end{aligned}$$

Because $Y_0 = 1$, induction gives $\mathbb{E}[Y_t] \leq 1$ for every finite t. Thus $(Y_t)_{t\geq 0}$ is an integrable nonnegative supermartingale.

Step 3. Bound the probability of crossing the cutoff.

Since $p > 0$, crossing B in norm is the same as crossing B^{2p} for Y_t. Ville's maximal inequality (Lemma F.4) gives

$$\begin{aligned}\mathbb{P}\{\exists\, t \geq 0 : \|\boldsymbol{x}_t\| > B\} &= \mathbb{P}\{\exists\, t \geq 0 : Y_t > B^{2p}\} \\ &\leq \mathbb{P}\Big\{\sup_{t\geq 0} Y_t \geq B^{2p}\Big\} \\ &\leq \frac{\mathbb{E}[Y_0]}{B^{2p}} && \text{Ville's maximal inequality} \\ &= B^{-2p} && Y_0 = \|\boldsymbol{x}_0\|^{2p} = 1 \\ &= B^{-(a-h^2)/h^2}. && \text{the definition of } p\end{aligned}$$

If $a \geq 2h^2$, then

$$\begin{aligned}\frac{a-h^2}{h^2} &\geq \frac{a}{2h^2} \\ &= \frac{\eta_{\mathrm{p}}(\lambda-\lambda_\star)}{2\eta_{\mathrm{p}}^2\sigma^2} = \frac{\lambda-\lambda_\star}{2\eta_{\mathrm{p}}\sigma^2}.\end{aligned}$$

Since $B > 1$, the probability bound is therefore at most $B^{-(\lambda-\lambda_\star)/(2\eta_{\mathrm{p}}\sigma^2)}$. Finally, recall $g_+ = 2\eta_{\mathrm{p}}\sigma^2 \max\{1, \log(N/\delta)/\log B\}$. If $\lambda - \lambda_\star \geq g_+$, its two defining lower bounds give

$$\begin{aligned}a = \eta_{\mathrm{p}}(\lambda-\lambda_\star) &\geq 2\eta_{\mathrm{p}}^2\sigma^2 = 2h^2, \\ \frac{\lambda-\lambda_\star}{2\eta_{\mathrm{p}}\sigma^2} &\geq \frac{\log(N/\delta)}{\log B}.\end{aligned}$$

Consequently,

$$\begin{aligned}B^{-(\lambda-\lambda_\star)/(2\eta_{\mathrm{p}}\sigma^2)} &\leq B^{-\log(N/\delta)/\log B} \\ &= \exp\big(-\log(N/\delta)\big) = \frac{\delta}{N}.\end{aligned}$$

This completes the proof of Lemma F.2. □

F.4 Auxiliary Results

We state the three standard concentration results used above. Throughout, $(\mathcal{F}_t)_{t\geq 0}$ is a filtration.

Lemma F.3 (Hoeffding's lemma, conditional form; Hoeffding (1963), see also Lemma 2.2 of Boucheron et al. (2013))**.** *Let D be a random variable and $\mathcal{G}$ a sigma-algebra such that $\mathbb{E}[D \mid \mathcal{G}] = 0$ and $|D| \leq a$ almost surely for a constant $a \geq 0$. Then for every $\theta \in \mathbb{R}$,*

$$\mathbb{E}\big[e^{\theta D} \mid \mathcal{G}\big] \leq e^{\theta^2 a^2/2}.$$

Lemma F.4 (Ville's maximal inequality for nonnegative supermartingales; Ville (1939), Theorem 4.8.4 of Durrett (2019))**.** *Let $(Y_t)_{t\geq 0}$ be a nonnegative supermartingale with respect to $(\mathcal{F}_t)_{t\geq 0}$. Then for every $c > 0$,*

$$\mathbb{P}\Big\{\sup_{t\geq 0} Y_t \geq c\Big\} \leq \frac{\mathbb{E}[Y_0]}{c}.$$

Lemma F.5 (Azuma–Hoeffding inequality; Azuma (1967); Hoeffding (1963))**.** *Let $(D_t)_{t\geq 0}$ be random variables such that D_t is $\mathcal{F}_{t+1}$-measurable, $\mathbb{E}[D_t \mid \mathcal{F}_t] = 0$, and $|D_t| \leq a_t$ almost surely for constants $a_t \geq 0$. Let $S_t := \sum_{s<t} D_s$. For every integer $T_{\mathrm{p}} \geq 1$ with $\sum_{t<T_{\mathrm{p}}} a_t^2 > 0$ and every $r > 0$,*

$$\mathbb{P}\{S_{T_{\mathrm{p}}} \leq -r\} \leq \exp\Big(-\frac{r^2}{2\sum_{t<T_{\mathrm{p}}} a_t^2}\Big).$$

G Proof of the Compensated Finite-Horizon Threshold Bound

The proof compares the projected critical LR, the compensated critical LR, and the finite-horizon detector output. Appendix G.1 gives the stability results used in this comparison. Appendix G.2 then proves the main bound from one compensation bound and one horizon bound. We prove these two bounds in Appendices G.3 and G.4. The proofs of the stability results and the auxiliary matrix facts follow in Appendices G.5 and G.6.

G.1 Stability Criterion

We first analyze a generic momentum recurrence. Let the mean be a symmetric positive definite $\boldsymbol{C} \in \mathbb{S}_n$, let the parameters satisfy $-1 < b < 1$ and $\alpha > 0$, and let the noise matrices be i.i.d. and symmetric $\boldsymbol{\Delta}_t$ with $\mathbb{E}\boldsymbol{\Delta}_t = 0$, $\mathbb{E}\,\|\boldsymbol{\Delta}_t\|_F^2 < \infty$, independent of the deterministic initial pair $(\boldsymbol{x}_0, \boldsymbol{x}_{-1})$:

$$\boldsymbol{x}_{t+1} = \big[(1+b)\boldsymbol{I} - \alpha(\boldsymbol{C} + \boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t - b\,\boldsymbol{x}_{t-1}. \tag{61}$$

The history state $\boldsymbol{y}_t := (\boldsymbol{x}_t, \boldsymbol{x}_{t-1})^\top$ evolves by $\boldsymbol{y}_{t+1} = \boldsymbol{F}_t \boldsymbol{y}_t$ with

$$\boldsymbol{F}_t := \begin{pmatrix} (1+b)\boldsymbol{I} - \alpha(\boldsymbol{C} + \boldsymbol{\Delta}_t) & -b\,\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{pmatrix}, \qquad \mathcal{T}(\boldsymbol{S}) := \mathbb{E}[\boldsymbol{F}_t \boldsymbol{S} \boldsymbol{F}_t^\top], \tag{62}$$

where $\mathcal{T}$, the *second-moment map*, sends the second moment of $\boldsymbol{y}_t$ to that of $\boldsymbol{y}_{t+1}$. Strict mean square stability means that $\mathbb{E}\|\boldsymbol{y}_t\|^2 \to 0$ for every deterministic initial pair.

We write $\mathbb{S}_n$ and $\mathbb{S}_n^+$ for the symmetric and positive semidefinite $n \times n$ matrices, respectively, and $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \operatorname{tr}(\boldsymbol{X}\boldsymbol{Y})$ for their Frobenius inner product. The largest eigenvalue of the generic mean is $L_{\boldsymbol{C}} := \lambda_{\max}(\boldsymbol{C})$. This distinguishes it from the paper's $L = \lambda_{\max}(\boldsymbol{A})$.

The stability quotient q gives both an exact test for stability and a quantitative bound when the moment spectral radius may exceed one. The proofs are in Appendix G.5.

Recall the quotient and its maximum from (49)–(50):

$$q_{\boldsymbol{C},b}(\boldsymbol{X}) = \frac{\mathbb{E}\operatorname{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) + c_b \operatorname{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) - k_b\,\|[\boldsymbol{C}, \boldsymbol{X}]\|_F^2}{2\operatorname{tr}(\boldsymbol{C}\boldsymbol{X}^2)},$$

$$r_{\boldsymbol{C}}(b) = \max_{\boldsymbol{X} \succeq 0,\ \|\boldsymbol{X}\|_F = 1} q_{\boldsymbol{C},b}(\boldsymbol{X}), \qquad c_b = \frac{1-b}{1+b}, \quad k_b = \frac{b}{1-b^2}.$$

The first result compares the normalized inverse coefficient $(1-b)/\alpha$ with the largest value of this quotient.

Theorem G.1 (Exact criterion). *For the momentum recurrence (61) with $\boldsymbol{C} \succ 0$, $-1 < b < 1$, $\alpha > 0$, and i.i.d. symmetric centered noise of finite second moment, strict mean-square stability holds if and only if*

$$\frac{1-b}{\alpha} > r_{\boldsymbol{C}}(b). \tag{63}$$

Moreover, $r_{\boldsymbol{C}}(b) \geq c_b L_{\boldsymbol{C}}/2 > 0$. For fixed b, the stable coefficients form the interval $0 < \alpha < (1-b)/r_{\boldsymbol{C}}(b)$.

Proposition G.2 (A spectral-radius bound controls the stability quotient). *For the same recurrence, suppose additionally that $0 \leq b < 1$. Let $R \geq 1$ satisfy $\rho(\mathcal{T}) \leq R$. For every $\boldsymbol{X} \succeq 0$ with $\|\boldsymbol{X}\|_F = 1$, set $\mu := \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}^2) > 0$. Then*

$$q_{\boldsymbol{C},b}(\boldsymbol{X}) - \frac{1-b}{\alpha} \leq (R-1)\left[\frac{2b}{(R+b)\alpha} + \frac{(1-b/R)^2}{2\mu\alpha^2}\right]. \tag{64}$$

The test matrix can be chosen freely and need not maximize q. At $R = 1$, the result gives $q_{\boldsymbol{C},b}(\boldsymbol{X}) \leq (1-b)/\alpha$. For $R > 1$, it controls the amount by which this inequality can fail. When $b = 0$, the bound reduces to $(R-1)/(2\mu\alpha^2)$. In Appendix G.4, the detector supplies $R = \rho_T$, and the critical test matrix converts the left side into a bound on the inverse LR.

Changing the momentum. First compare two momentum values while keeping the mean, the noise and the test matrix fixed. Recall $c_b = (1-b)/(1+b)$ and $k_b = b/(1-b^2)$. For $-1 < b' \le b < 1$, we have $c_{b'} \ge c_b$ and $k_{b'} \le k_b$. The noise term cancels when we subtract the two quotients, so for every nonzero PSD $\boldsymbol{X}$,

$$\begin{aligned} q_{\boldsymbol{C},b'}(\boldsymbol{X}) - q_{\boldsymbol{C},b}(\boldsymbol{X}) &= \frac{(c_{b'} - c_b)\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) + (k_b - k_{b'})\,\|[\boldsymbol{C},\boldsymbol{X}]\|_F^2}{2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)} \\ &\ge 0. \end{aligned} \tag{65}$$

Both coefficients and both quadratic terms in the numerator are nonnegative. Thus $q_{\boldsymbol{C},b'}(\boldsymbol{X}) \ge q_{\boldsymbol{C},b}(\boldsymbol{X})$ for every admissible test matrix $\boldsymbol{X}$.

Shifting the mean. A scalar shift changes only the mean-dependent terms of the quotient. The following identity will be used in both the quantitative stability bound and the compensation comparison. Its proof is in Appendix G.5.2.

Lemma G.3 (The shift identity)**.** *Let $\boldsymbol{C} \succ 0$ on $\mathbb{R}^n$ with noise $\boldsymbol{\Delta}_t$ and momentum $b \in (-1,1)$, and let $h \ge 0$. The shifted mean $\boldsymbol{C} + h\boldsymbol{I}$ carries the same noise. For every $\boldsymbol{X} \in \mathbb{S}_n^+$ with $\|\boldsymbol{X}\|_F = 1$ and $\mu := \mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)$,*

$$q_{\boldsymbol{C}+h\boldsymbol{I},b}(\boldsymbol{X}) = \frac{\mu\, q_{\boldsymbol{C},b}(\boldsymbol{X}) + c_b h(\mu + h/2)}{\mu + h}. \tag{66}$$

Application to AGMs. Consider two affine runs $(\boldsymbol{\epsilon}_t^{(1)}, \boldsymbol{m}_t^{(1)})$ and $(\boldsymbol{\epsilon}_t^{(2)}, \boldsymbol{m}_t^{(2)})$ driven by the same mini-batches. Define their preconditioned differences by

$$\boldsymbol{x}_t := \boldsymbol{P}^{1/2}(\boldsymbol{\epsilon}_t^{(1)} - \boldsymbol{\epsilon}_t^{(2)}), \qquad \boldsymbol{n}_t := \boldsymbol{P}^{-1/2}(\boldsymbol{m}_t^{(1)} - \boldsymbol{m}_t^{(2)}).$$

Subtracting the two copies of (47) cancels the common batch gradient and checkpoint drift. Thus

$$\begin{aligned} \boldsymbol{n}_{t+1} &= \beta_1 \boldsymbol{n}_t + (1-\beta_1)(\boldsymbol{A}_{\mathsf{P}} - \lambda_{\mathrm{WD}}\boldsymbol{I} + \boldsymbol{\Delta}_t)\boldsymbol{x}_t, \\ \boldsymbol{x}_{t+1} &= (1 - \eta\lambda_{\mathrm{WD}})\boldsymbol{x}_t - \eta\boldsymbol{n}_{t+1}. \end{aligned} \tag{67}$$

The curvature term follows from $\boldsymbol{P}^{-1/2}(\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{P})\boldsymbol{P}^{-1/2} = \boldsymbol{A}_{\mathcal{B}_t} - \lambda_{\mathrm{WD}}\boldsymbol{I} + \lambda_{\mathsf{P}}\boldsymbol{I}$. These are the homogeneous difference dynamics of the compensated augmented system in (33). Its state $(\eta\boldsymbol{n}_t, \boldsymbol{x}_t)$ is written here in the coordinates $(\boldsymbol{x}_t, \boldsymbol{n}_t)$. To obtain a recurrence for the positions, set

$$\alpha := \eta(1-\beta_1), \qquad b := \beta_1(1 - \eta\lambda_{\mathrm{WD}}).$$

For $t \ge 1$, the preceding position update gives $-\eta\boldsymbol{n}_t = \boldsymbol{x}_t - (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{x}_{t-1}$. Substituting this into (67),

$$\begin{aligned} \boldsymbol{x}_{t+1} &= (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{x}_t - \eta\beta_1\boldsymbol{n}_t - \alpha(\boldsymbol{A}_{\mathsf{P}} - \lambda_{\mathrm{WD}}\boldsymbol{I} + \boldsymbol{\Delta}_t)\boldsymbol{x}_t \\ &= (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{x}_t + \beta_1\big[\boldsymbol{x}_t - (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{x}_{t-1}\big] - \alpha(\boldsymbol{A}_{\mathsf{P}} + \boldsymbol{\Delta}_t)\boldsymbol{x}_t + \alpha\lambda_{\mathrm{WD}}\boldsymbol{x}_t \\ &= \big[(1 + \beta_1 - \eta\lambda_{\mathrm{WD}} + \alpha\lambda_{\mathrm{WD}})\boldsymbol{I} - \alpha(\boldsymbol{A}_{\mathsf{P}} + \boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t - b\boldsymbol{x}_{t-1}. \end{aligned}$$

The scalar coefficient simplifies because

$$1 + \beta_1 - \eta\lambda_{\mathrm{WD}} + \alpha\lambda_{\mathrm{WD}} = 1 + \beta_1 - \eta\lambda_{\mathrm{WD}} + \eta(1-\beta_1)\lambda_{\mathrm{WD}} = 1 + \beta_1(1 - \eta\lambda_{\mathrm{WD}}) = 1 + b.$$

Thus the positions satisfy

$$\boldsymbol{x}_{t+1} = \big[(1+b)\boldsymbol{I} - \alpha(\boldsymbol{A}_{\mathsf{P}} + \boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t - b\,\boldsymbol{x}_{t-1}. \tag{68}$$

The effective curvature already includes weight decay. Its remaining contribution is absorbed into the *effective momentum* b. The LR enters through α and b. The elimination is valid for every η, including $\eta\lambda_{\mathrm{WD}} = 1$, where $b = 0$ and the recurrence has first order.

The generic criterion applies to two positive definite means. The compensated system (68) uses $\boldsymbol{C} = \boldsymbol{A}_{\mathsf{P}}$ and noise $\boldsymbol{\Delta}_t$ in dimension d. The projected system uses $\boldsymbol{C} = \boldsymbol{A}^+$ and noise $\boldsymbol{\Delta}_t^+$ in the retained dimension p. It is the homogeneous difference recurrence associated with the projected affine system in Definition B.2. In both systems, $(b, \alpha) = (\beta_1(1-\eta\lambda_{\mathrm{WD}}), \eta(1-\beta_1))$, so that

$$\frac{1-b}{\alpha} = \frac{1 - \beta_1 + \beta_1\eta\lambda_{\mathrm{WD}}}{\eta(1-\beta_1)} = \frac{1}{\eta} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}. \tag{69}$$

The normalized inverse LR $(1-b)/\alpha$ and $1/\eta$ differ by a constant.

For the two AGM instances, we write $\mathcal{T}_\eta$ for the projected system and $\mathcal{T}_\eta^{\mathsf{P}}$ for the compensated system. For either instance, this map has the same spectral radius as the corresponding second-moment operator for the augmented position–momentum state in Appendix D.1: the two states are related by an invertible linear change when $\eta\lambda_{\mathrm{WD}} \neq 1$, and both spectral radii depend continuously on η (Appendix G.6).

We now vary the LR along the AGM parameter curve, keeping $\boldsymbol{C}$, the noise law, β_1 and λ_{WD} fixed. Here $0 \le \beta_1 < 1$ and $\lambda_{\mathrm{WD}} \ge 0$. Let $\eta_{\max} := (1+\beta_1)/(\beta_1\lambda_{\mathrm{WD}})$ when $\beta_1\lambda_{\mathrm{WD}} > 0$, and $\eta_{\max} := \infty$ otherwise. This is the upper endpoint of the range where $b > -1$. Define

$$G_{\boldsymbol{C}}(\eta) := \frac{1}{\eta} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1} - r_{\boldsymbol{C}}\big(\beta_1(1-\eta\lambda_{\mathrm{WD}})\big), \qquad 0 < \eta < \eta_{\max}. \tag{70}$$

The next corollary shows that this function crosses zero once, so the fixed-parameter criterion also gives a unique critical LR along the AGM parameter curve.

Corollary G.4 (AGM critical LR). *For an AGM instance of (61) with $\boldsymbol{C} \succ 0$, the centered i.i.d. symmetric noise of finite second moment, and the parameter curve above, $G_{\boldsymbol{C}}$ has a unique root η_{crit}. The strictly mean-square stable LRs are $(0, \eta_{\mathrm{crit}})$. The second-moment spectral radius equals one at η_{crit} and is strictly greater than one for every $\eta > \eta_{\mathrm{crit}}$. For $0 < \eta_2 < \eta_1 < \eta_{\max}$,*

$$G_{\boldsymbol{C}}(\eta_2) - G_{\boldsymbol{C}}(\eta_1) \ge \frac{1}{\eta_2} - \frac{1}{\eta_1}. \tag{71}$$

The corollary gives the projected and compensated thresholds $\eta_\star$ and $\eta_\star^{\mathsf{P}}$ from Definition D.13, with critical effective momenta in $(-1, 1)$. These agree with the corresponding thresholds defined by bounded affine moments by Proposition B.3 and its argument for $\boldsymbol{A}_{\mathsf{P}} \succ 0$.

At the unique root, write $b_{\mathrm{crit}} := \beta_1(1-\eta_{\mathrm{crit}}\lambda_{\mathrm{WD}})$. The normalized inverse rate equals the quotient maximum:

$$r_{\boldsymbol{C}}(b_{\mathrm{crit}}) = \frac{1-b_{\mathrm{crit}}}{(1-\beta_1)\eta_{\mathrm{crit}}} = \frac{1}{\eta_{\mathrm{crit}}} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}. \tag{72}$$

The offset $\beta_1\lambda_{\mathrm{WD}}/(1-\beta_1)$ is the same for the projected and compensated systems.

Relation to the exact SGD threshold. When $b = 0$, we have $c_0 = 1$ and $k_0 = 0$. Centering of the noise gives

$$q_{\boldsymbol{C},0}(\boldsymbol{X}) = \frac{\mathbb{E}\,\mathrm{tr}\big(\boldsymbol{X}(\boldsymbol{C}+\boldsymbol{\Delta}_t)\boldsymbol{X}(\boldsymbol{C}+\boldsymbol{\Delta}_t)\big)}{2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)}.$$

Under the sampling and positive-semidefinite sample-Hessian assumptions of Mulayoff & Michaeli (2024, Theorem 5), maximizing this quotient recovers their exact SGD threshold on the positive-curvature subspace. Our criterion extends it to momentum and allows indefinite curvature samples with a positive definite mean. For the normalized momentum update used here, $\alpha = \eta(1-\beta_1)$. Thus, without weight decay,

$$\eta_{\mathrm{crit}} = \frac{1}{r_{\boldsymbol{C}}(0)} \quad \text{for SGD}, \qquad \eta_{\mathrm{crit}} = \frac{1}{r_{\boldsymbol{C}}(\beta_1)} \quad \text{with momentum}.$$

Momentum changes the deterministic-curvature coefficient and adds a commutator term, so it does not universally rescale the SGD threshold. When $\beta_1\lambda_{\mathrm{WD}} > 0$, the threshold is the unique root of $G_{\boldsymbol{C}}$ in (70).

The noise-free threshold. Set $\boldsymbol{\Delta}_t = 0$ while retaining the same mean $\boldsymbol{C}$, whose largest eigenvalue is $L_{\boldsymbol{C}}$. The noise term in $q_{\boldsymbol{C},b}$ then vanishes. The remaining numerator is at most $c_b L_{\boldsymbol{C}}\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)$ by (105), so division by its denominator gives $q_{\boldsymbol{C},b}(\boldsymbol{X}) \le c_b L_{\boldsymbol{C}}/2$ for every admissible $\boldsymbol{X}$. Equality holds at $\boldsymbol{X} = \boldsymbol{u}\boldsymbol{u}^\top$ for a unit top eigenvector of $\boldsymbol{C}$. Together with the rank-one lower bound,

$$r_{\boldsymbol{C}}(b) = \frac{c_b L_{\boldsymbol{C}}}{2} \text{ without noise}, \qquad r_{\boldsymbol{C}}(b) \ge \frac{c_b L_{\boldsymbol{C}}}{2} \text{ for arbitrary noise}. \tag{73}$$

Thus noise cannot increase the critical LR. For the noise-free AGM system, the equality in the stability criterion becomes

$$\alpha L_{\boldsymbol{C}} = 2(1+b),$$
$$\eta(1-\beta_1)L_{\boldsymbol{C}} = 2(1+\beta_1) - 2\beta_1\lambda_{\mathrm{WD}}\eta.$$

Solving for the LR gives

$$\eta_{\mathrm{det}} = \frac{2(1+\beta_1)}{(1-\beta_1)L_{\boldsymbol{C}} + 2\beta_1\lambda_{\mathrm{WD}}}, \tag{74}$$

the noise-free critical LR of this AGM system.

G.2 Proof of the Threshold Bound

The detector runs the compensated system, whereas EoS is defined by the projected system. We therefore compare three LRs: the detector output $\widehat{\eta}_\star(M,T)$, the compensated critical rate $\eta_\star^{\mathsf{P}}$, and the projected critical rate $\eta_\star$. Their inverse difference splits into two terms:

$$\frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star} = \underbrace{\left(\frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star^{\mathsf{P}}}\right)}_{\text{finite-horizon error}} + \underbrace{\left(\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star}\right)}_{\text{compensation error}}. \tag{75}$$

The first comes from observing a run for only T steps. The second remains even with an unlimited horizon, because compensation changes the system being tested. The theorem bounds both terms without assuming that the Hessian noise is small.

Proof from the two error bounds. The following propositions control the two terms in (75). They are proved in Appendices G.3 and G.4, respectively. Recall that $\mu_\star$ is the curvature at a test matrix maximizing the projected stability quotient. The function Φ gives the scalar-shift correction, and δ_{cov} is the maximum normalized noise mismatch between the projected and compensated systems. Their formulas are given in (52), (54), and (53), respectively.

Proposition G.5 (Compensation error)**.** *Under Assumption D.14,*

$$\min\{\Phi(\mu_\star), 0\} \;\le\; \frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star} \;\le\; \max\{\Phi(L) + \delta_{\mathrm{cov}}, 0\}. \tag{76}$$

If $\beta_1\lambda_{\mathrm{WD}} = 0$, the clipping can be omitted: $\Phi(\mu_\star) \le 1/\eta_\star^{\mathsf{P}} - 1/\eta_\star \le \Phi(L) + \delta_{\mathrm{cov}}$.

Proposition G.6 (Finite-horizon error)**.** *Under Assumption D.14, let $\eta \in \mathcal{I}$ be accepted by the detector. If $\eta \le \eta_\star^{\mathsf{P}}$, then $1/\eta \ge 1/\eta_\star^{\mathsf{P}}$. Otherwise $b_\star^{\mathsf{P}} \ge 0$ and*

$$\frac{1}{\eta} \ge \frac{1}{\eta_\star^{\mathsf{P}}} - H_T, \tag{77}$$

with H_T given by (56). In particular, (77) holds for every accepted LR whenever $\eta_\star^{\mathsf{P}} \in \mathcal{I}$.

Proof of Theorem D.16. We bound the two inverse-rate differences in (75) separately and then add them. Recall that $\widehat{\eta}_\star(M,T)$ is the supremum of the accepted rates in the search window. By Assumption D.15, $\eta_\star^{\mathsf{P}}$ belongs to that window and is accepted, so $\widehat{\eta}_\star(M,T) \ge \eta_\star^{\mathsf{P}}$.

By the finite-horizon bound (Proposition G.6), every accepted rate satisfies $1/\eta \ge 1/\eta_\star^{\mathsf{P}} - H_T$. We now pass this inequality to the supremum without assuming that the accepted rates form an interval. Every accepted η is at most $\widehat{\eta}_\star(M,T)$, so its reciprocal is at least $1/\widehat{\eta}_\star(M,T)$. Conversely, by the definition of a supremum, there is a sequence of accepted rates $\eta_j \to \widehat{\eta}_\star(M,T)$. The window is bounded and bounded away from zero, so

$$\inf_{\substack{\eta\in\mathcal{I}\\ \eta \text{ accepted}}} \frac{1}{\eta} \ge \frac{1}{\widehat{\eta}_\star(M,T)},$$

$$\inf_{\substack{\eta\in\mathcal{I}\\ \eta \text{ accepted}}} \frac{1}{\eta} \le \lim_{j\to\infty} \frac{1}{\eta_j} = \frac{1}{\widehat{\eta}_\star(M,T)}.$$

The two inequalities identify the infimum. Applying the bound for each accepted rate and then using $\widehat{\eta}_\star(M,T) \ge \eta_\star^{\mathsf{P}}$ gives

$$\frac{1}{\widehat{\eta}_\star(M,T)} = \inf_{\substack{\eta\in\mathcal{I}\\ \eta \text{ accepted}}} \frac{1}{\eta} \ge \frac{1}{\eta_\star^{\mathsf{P}}} - H_T,$$

$$-H_T \le \frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star^{\mathsf{P}}} \le 0.$$

The compensation bound (Proposition G.5) controls the remaining difference. Thus

$$\frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star} = \left(\frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star^{\mathsf{P}}}\right) + \left(\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star}\right),$$

$$\min\{\Phi(\mu_\star), 0\} - H_T \le \frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star} \le \max\{\Phi(L) + \delta_{\mathrm{cov}}, 0\}.$$

When $\beta_1\lambda_{\mathrm{WD}} = 0$, the compensation bound has no clipping, so neither does the sum. When $\beta_1 = 0$, the definitions give $b_\star^{\mathsf{P}} = 0$ and $\alpha_\star^{\mathsf{P}} = \eta_\star^{\mathsf{P}}$. The first term in the bracket defining H_T vanishes, and the second gives $H_T = (\rho_T - 1)/(2\mu_\star^{\mathsf{P}}(\eta_\star^{\mathsf{P}})^2)$. Finally, with $x := (2/T)\log(M(1+\beta_1)/(1-\beta_1))$,

$$\rho_T - 1 = e^x - 1 \le xe^x = O\big((1+\log M)/T\big) \quad \text{as } (1+\log M)/T \to 0.$$

The search window gives $0 \le b_\star^{\mathsf{P}} < 1$, and $\rho_T \ge 1$. Thus the bracket in (56) has the fixed upper bound

$$\frac{2b_\star^{\mathsf{P}}}{(\rho_T + b_\star^{\mathsf{P}})\alpha_\star^{\mathsf{P}}} + \frac{(1 - b_\star^{\mathsf{P}}/\rho_T)^2}{2\mu_\star^{\mathsf{P}}(\alpha_\star^{\mathsf{P}})^2} \le \frac{2b_\star^{\mathsf{P}}}{(1+b_\star^{\mathsf{P}})\alpha_\star^{\mathsf{P}}} + \frac{1}{2\mu_\star^{\mathsf{P}}(\alpha_\star^{\mathsf{P}})^2}.$$

The denominator in each term is positive and fixed with the model. Multiplying by the preceding bound on $\rho_T - 1$ gives the stated order of H_T and completes the proof of Theorem D.16. □

G.3 Compensation Error

This section proves Proposition G.5 by comparing the two systems at the same momentum and the same test matrix. A scalar shift of the mean gives an exact identity. Adding the discarded directions does not change the threshold if their noise is removed, and restoring that noise costs at most δ_{cov}.

Recall that $\boldsymbol{A} = \boldsymbol{D}_{t_0} + \lambda_{\mathrm{WD}}\boldsymbol{I}$ is the effective curvature and $\boldsymbol{\Delta}_t = \boldsymbol{A}_{\mathcal{B}_t} - \boldsymbol{A}$ is its centered noise. Its retained block is $\boldsymbol{A}^+ \succ 0$, its compensated mean is $\boldsymbol{A}_{\mathsf{P}} = \boldsymbol{A} + \lambda_{\mathsf{P}}\boldsymbol{I} \succ 0$, and $L = \lambda_{\max}(\boldsymbol{A}^+)$. The quotient and its maximum from (49)–(50) are

$$q_{\boldsymbol{C},b}(\boldsymbol{X}) = \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) + c_b\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) - k_b\,\|[\boldsymbol{C},\boldsymbol{X}]\|_F^2}{2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)},$$

$$r_{\boldsymbol{C}}(b) = \max_{\boldsymbol{X}\succeq 0,\ \|\boldsymbol{X}\|_F = 1} q_{\boldsymbol{C},b}(\boldsymbol{X}), \qquad c_b = \frac{1-b}{1+b}, \quad k_b = \frac{b}{1-b^2}.$$

The noise in this formula is $\boldsymbol{\Delta}_t^+$ for the projected mean $\boldsymbol{A}^+$ and the full $\boldsymbol{\Delta}_t$ for the compensated mean $\boldsymbol{A}_{\mathsf{P}}$. The shift identity (Lemma G.3) was established in Appendix G.1 and applies here with the corresponding noise law fixed.

Proof of Proposition G.5. We compare first at $b_\star$, then at the systems' own critical momenta using (65) and its consequence for $r_{\boldsymbol{C}}$ proved in Appendix G.5.3.

Recall that $b_\star = \beta_1(1 - \eta_\star\lambda_{\mathrm{WD}})$ is the projected critical momentum and $c_\star = (1-b_\star)/(1+b_\star)$. The matrix $\boldsymbol{X}_\star$ maximizes $q_{\boldsymbol{A}^+,b_\star}$, with $\|\boldsymbol{X}_\star\|_F = 1$ and $\mu_\star = \mathrm{tr}(\boldsymbol{A}^+\boldsymbol{X}_\star^2)$. All threshold functions in Steps 1–2 are evaluated at the same momentum $b_\star$.

Step 1. Shift the mean on the retained subspace.

Fix $\boldsymbol{X} \succeq 0$ with $\|\boldsymbol{X}\|_F = 1$ in the retained subspace, and write $\mu := \mathrm{tr}(\boldsymbol{A}^+\boldsymbol{X}^2)$, as in the shift identity. Since $0 \prec \boldsymbol{A}^+ \preceq L\boldsymbol{I}$ and $\boldsymbol{X}^2 \succeq 0$,

$$0 < \mu = \mathrm{tr}(\boldsymbol{A}^+\boldsymbol{X}^2) \le L\,\mathrm{tr}(\boldsymbol{X}^2) = L.$$

Apply (66) with $\boldsymbol{C} = \boldsymbol{A}^+$, $b = b_\star$ and $h = \lambda_{\mathsf{P}}$. Subtracting $r_{\boldsymbol{A}^+}(b_\star)$ and collecting its coefficients gives

$$\begin{aligned} &q_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I},b_\star}(\boldsymbol{X}) - r_{\boldsymbol{A}^+}(b_\star) \\ &\quad= \frac{\mu\, q_{\boldsymbol{A}^+,b_\star}(\boldsymbol{X}) + c_\star\lambda_{\mathsf{P}}(\mu + \lambda_{\mathsf{P}}/2)}{\mu + \lambda_{\mathsf{P}}} - r_{\boldsymbol{A}^+}(b_\star) \\ &\quad= \frac{\mu\big(q_{\boldsymbol{A}^+,b_\star}(\boldsymbol{X}) - r_{\boldsymbol{A}^+}(b_\star)\big)}{\mu + \lambda_{\mathsf{P}}} + \frac{\lambda_{\mathsf{P}}\big[c_\star(\mu + \lambda_{\mathsf{P}}/2) - r_{\boldsymbol{A}^+}(b_\star)\big]}{\mu + \lambda_{\mathsf{P}}}. \end{aligned}$$

Recall that Φ is the compensation term defined in (54). At the curvature $\mu = \mathrm{tr}(\boldsymbol{A}^+\boldsymbol{X}^2)$, its definition gives

$$\Phi(\mu) = \frac{c_\star \lambda_{\mathsf{P}}}{2} - \frac{\lambda_{\mathsf{P}}}{\mu + \lambda_{\mathsf{P}}}\left(r_{\boldsymbol{A}^+}(b_\star) - \frac{c_\star \mu}{2}\right), \qquad c_\star = \frac{1 - b_\star}{1 + b_\star}.$$

To identify the second term above with $\Phi(\mu)$, expand its numerator as follows:

$$\begin{aligned}
&\frac{\lambda_{\mathsf{P}}\left[c_\star(\mu + \lambda_{\mathsf{P}}/2) - r_{\boldsymbol{A}^+}(b_\star)\right]}{\mu + \lambda_{\mathsf{P}}} \\
&\quad = \frac{\lambda_{\mathsf{P}}\left[c_\star(\mu + \lambda_{\mathsf{P}})/2 + c_\star\mu/2 - r_{\boldsymbol{A}^+}(b_\star)\right]}{\mu + \lambda_{\mathsf{P}}} \\
&\quad = \frac{c_\star \lambda_{\mathsf{P}}}{2} - \frac{\lambda_{\mathsf{P}}}{\mu + \lambda_{\mathsf{P}}}\left(r_{\boldsymbol{A}^+}(b_\star) - \frac{c_\star \mu}{2}\right) \\
&\quad = \Phi(\mu).
\end{aligned}$$

Consequently,

$$\begin{aligned}
&q_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, b_\star}(\boldsymbol{X}) - r_{\boldsymbol{A}^+}(b_\star) \\
&\quad = \frac{\mu}{\mu + \lambda_{\mathsf{P}}}\left(q_{\boldsymbol{A}^+, b_\star}(\boldsymbol{X}) - r_{\boldsymbol{A}^+}(b_\star)\right) + \Phi(\mu).
\end{aligned} \tag{78}$$

For $0 \le s \le L$, the rank-one bound (108) yields

$$r_{\boldsymbol{A}^+}(b_\star) - \frac{c_\star s}{2} \ge \frac{c_\star(L - s)}{2} \ge 0, \qquad \text{hence} \qquad \Phi(s) \le \frac{c_\star \lambda_{\mathsf{P}}}{2}.$$

To compare $\Phi(\mu)$ with $\Phi(L)$, differentiate its equivalent form $\Phi(s) = \lambda_{\mathsf{P}}[c_\star(s + \lambda_{\mathsf{P}}/2) - r_{\boldsymbol{A}^+}(b_\star)]/(s + \lambda_{\mathsf{P}})$:

$$\begin{aligned}
\Phi'(s) &= \lambda_{\mathsf{P}} \frac{c_\star(s + \lambda_{\mathsf{P}}) - \left[c_\star(s + \lambda_{\mathsf{P}}/2) - r_{\boldsymbol{A}^+}(b_\star)\right]}{(s + \lambda_{\mathsf{P}})^2} \\
&= \frac{\lambda_{\mathsf{P}}\left[r_{\boldsymbol{A}^+}(b_\star) + c_\star \lambda_{\mathsf{P}}/2\right]}{(s + \lambda_{\mathsf{P}})^2} > 0.
\end{aligned}$$

The numerator is positive because $\lambda_{\mathsf{P}} > 0$, $c_\star > 0$, and $r_{\boldsymbol{A}^+}(b_\star) \ge c_\star L/2 > 0$.

Since $q_{\boldsymbol{A}^+, b_\star}(\boldsymbol{X}) \le r_{\boldsymbol{A}^+}(b_\star)$, the first term in (78) is nonpositive. Moreover, Φ is increasing and $\mu \le L$. Thus, for every admissible $\boldsymbol{X}$,

$$\begin{aligned}
q_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, b_\star}(\boldsymbol{X}) - r_{\boldsymbol{A}^+}(b_\star) &\le \Phi(\mu) \\
&\le \Phi(L).
\end{aligned}$$

Writing out the corresponding maximum gives

$$\begin{aligned}
&r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) \\
&\quad = \max_{\boldsymbol{X} \in \mathbb{S}_p^+,\ \|\boldsymbol{X}\|_F = 1} \left[q_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, b_\star}(\boldsymbol{X}) - r_{\boldsymbol{A}^+}(b_\star)\right] \\
&\quad \le \max_{\boldsymbol{X} \in \mathbb{S}_p^+,\ \|\boldsymbol{X}\|_F = 1} \Phi\big(\mathrm{tr}(\boldsymbol{A}^+\boldsymbol{X}^2)\big) \\
&\quad \le \Phi(L).
\end{aligned}$$

The first equality moves the constant $r_{\boldsymbol{A}^+}(b_\star)$ inside the maximum. Each subsequent inequality uses a bound valid at every matrix in this same set.

For the lower bound, the maximizing matrix $\boldsymbol{X}_\star$ satisfies $q_{\boldsymbol{A}^+, b_\star}(\boldsymbol{X}_\star) = r_{\boldsymbol{A}^+}(b_\star)$. Thus

$$\begin{aligned}
&r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) \\
&\quad \ge q_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, b_\star}(\boldsymbol{X}_\star) - r_{\boldsymbol{A}^+}(b_\star) \\
&\quad = \frac{\mu_\star}{\mu_\star + \lambda_{\mathsf{P}}}\left[q_{\boldsymbol{A}^+, b_\star}(\boldsymbol{X}_\star) - r_{\boldsymbol{A}^+}(b_\star)\right] + \Phi(\mu_\star) \\
&\quad = \Phi(\mu_\star).
\end{aligned}$$

The inequality tests the shifted maximum at $\boldsymbol{X}_\star$, the next line uses (78), and the last line uses its maximizing property. We have proved

$$\Phi(\mu_\star) \le r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) \le \Phi(L).$$

Step 2. Restore the discarded directions and their noise.

Fix a retained-space matrix $\boldsymbol{X} \in \mathbb{S}_p^+$ with $\|\boldsymbol{X}\|_F = 1$. Its embedding $\operatorname{diag}(\boldsymbol{X}, 0)$ is positive semidefinite and has the same Frobenius norm, since

$$\operatorname{diag}(\boldsymbol{X}, 0)^2 = \operatorname{diag}(\boldsymbol{X}^2, 0),$$
$$\|\operatorname{diag}(\boldsymbol{X}, 0)\|_F^2 = \operatorname{tr}\big(\operatorname{diag}(\boldsymbol{X}^2, 0)\big) = \operatorname{tr}(\boldsymbol{X}^2) = 1.$$

We check the four terms of its quotient. For the noise term, block multiplication gives

$$\operatorname{diag}(\boldsymbol{X}, 0)\boldsymbol{\Delta}_t \operatorname{diag}(\boldsymbol{X}, 0) = \operatorname{diag}(\boldsymbol{X}\boldsymbol{\Delta}_t^+ \boldsymbol{X}, 0),$$
$$\operatorname{tr}\big(\operatorname{diag}(\boldsymbol{X}, 0)\boldsymbol{\Delta}_t \operatorname{diag}(\boldsymbol{X}, 0)\boldsymbol{\Delta}_t\big) = \operatorname{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t^+ \boldsymbol{X}\boldsymbol{\Delta}_t^+).$$

The equality holds for each sample, and therefore also after taking expectations. For the mean term and the denominator, the block diagonal form of $\boldsymbol{A}_{\mathsf{P}}$ gives

$$\operatorname{diag}(\boldsymbol{X}, 0)\boldsymbol{A}_{\mathsf{P}} = \operatorname{diag}\big(\boldsymbol{X}(\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}), 0\big),$$
$$\operatorname{tr}\big(\operatorname{diag}(\boldsymbol{X}, 0)\boldsymbol{A}_{\mathsf{P}} \operatorname{diag}(\boldsymbol{X}, 0)\boldsymbol{A}_{\mathsf{P}}\big) = \operatorname{tr}\big(\boldsymbol{X}(\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I})\boldsymbol{X}(\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I})\big),$$
$$\operatorname{tr}\big(\boldsymbol{A}_{\mathsf{P}} \operatorname{diag}(\boldsymbol{X}, 0)^2\big) = \operatorname{tr}\big((\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I})\boldsymbol{X}^2\big).$$

The commutator reduces to the retained block as well,

$$[\boldsymbol{A}_{\mathsf{P}}, \operatorname{diag}(\boldsymbol{X}, 0)] = \operatorname{diag}\big([\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, \boldsymbol{X}], 0\big),$$
$$\|[\boldsymbol{A}_{\mathsf{P}}, \operatorname{diag}(\boldsymbol{X}, 0)]\|_F^2 = \big\|[\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, \boldsymbol{X}]\big\|_F^2.$$

Substituting these equalities into the quotient proves

$$q_{\boldsymbol{A}_{\mathsf{P}}, b_\star}\big(\operatorname{diag}(\boldsymbol{X}, 0)\big) = q_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, b_\star}(\boldsymbol{X}). \tag{79}$$

The full maximization includes every embedded test matrix. In particular,

$$\begin{aligned} r_{\boldsymbol{A}_{\mathsf{P}}}(b_\star) &= \max_{\substack{\boldsymbol{X} \in \mathbb{S}_d^+ \\ \|\boldsymbol{X}\|_F = 1}} q_{\boldsymbol{A}_{\mathsf{P}}, b_\star}(\boldsymbol{X}) \\ &\ge \max_{\substack{\boldsymbol{X} \in \mathbb{S}_p^+ \\ \|\boldsymbol{X}\|_F = 1}} q_{\boldsymbol{A}_{\mathsf{P}}, b_\star}\big(\operatorname{diag}(\boldsymbol{X}, 0)\big) \\ &= \max_{\substack{\boldsymbol{X} \in \mathbb{S}_p^+ \\ \|\boldsymbol{X}\|_F = 1}} q_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, b_\star}(\boldsymbol{X}) \\ &= r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}}(b_\star). \end{aligned}$$

The inequality restricts the test matrices to embedded retained-space matrices, and the next equality uses (79).

For the upper bound, first remove the discarded and cross blocks of the noise. The resulting samples are

$$\boldsymbol{A}_{\mathsf{P}} + \Pi^+ \boldsymbol{\Delta}_t \Pi^+ = \begin{pmatrix} \boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I} + \boldsymbol{\Delta}_t^+ & 0 \\ 0 & \boldsymbol{A}^- + \lambda_{\mathsf{P}}\boldsymbol{I} \end{pmatrix}.$$

When $p < d$, this recurrence is strictly mean-square stable if and only if both block recurrences are stable. Sufficiency follows by adding their expected squared state norms, and necessity follows by initializing only one block. Both block means are positive definite because $\boldsymbol{A}_{\mathsf{P}} \succ 0$.

At the fixed momentum $b_\star$, let the coefficient $\alpha > 0$ vary. By the exact criterion (Theorem G.1) and the noise-free threshold (73), both blocks are strictly stable if and only if

$$\begin{cases} (1 - b_\star)/\alpha > r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}}(b_\star), \\ (1 - b_\star)/\alpha > c_\star\big(\lambda_{\max}(\boldsymbol{A}^-) + \lambda_{\mathsf{P}}\big)/2 \end{cases}$$

These two inequalities are equivalent to

$$\frac{1-b_\star}{\alpha} > \max\left\{ r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star), \frac{c_\star}{2}\big(\lambda_{\max}(\boldsymbol{A}^-) + \lambda_{\mathsf{P}}\big) \right\}.$$

Thus the maximum on the right is the threshold function of this block diagonal system. Since $\lambda_{\max}(\boldsymbol{A}^-) \le 0 < L$, the rank-one bound (108) gives

$$\begin{aligned}\frac{c_\star}{2}\big(\lambda_{\max}(\boldsymbol{A}^-) + \lambda_{\mathsf{P}}\big) &< \frac{c_\star}{2}(L + \lambda_{\mathsf{P}}) \\ &\le r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star).\end{aligned}$$

The threshold is therefore $r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star)$. When $p = d$, there is no discarded block and this equality holds directly.

Now fix a full-space test matrix $\boldsymbol{X} \in \mathbb{S}_d^+$ with $\|\boldsymbol{X}\|_F = 1$. Before the omitted noise is restored, its noise term is

$$\begin{aligned}&\mathbb{E}\,\mathrm{tr}\big(\boldsymbol{X}(\Pi^+ \boldsymbol{\Delta}_t \Pi^+)\boldsymbol{X}(\Pi^+ \boldsymbol{\Delta}_t \Pi^+)\big) \\ &\quad = \mathbb{E}\,\mathrm{tr}\big((\Pi^+ \boldsymbol{X} \Pi^+)\boldsymbol{\Delta}_t(\Pi^+ \boldsymbol{X} \Pi^+)\boldsymbol{\Delta}_t\big) \\ &\quad = \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^+ \boldsymbol{\Delta}_t^+ \boldsymbol{X}^+ \boldsymbol{\Delta}_t^+).\end{aligned}$$

The first equality uses cyclicity of the trace and $(\Pi^+)^2 = \Pi^+$, and the second follows by block multiplication. The mean, commutator and denominator are unchanged. Subtracting the change in the noise term therefore recovers the quotient before restoration:

$$\begin{aligned}&q_{\boldsymbol{A}_{\mathsf{P}}, b_\star}(\boldsymbol{X}) - \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) - \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^+\boldsymbol{\Delta}_t^+\boldsymbol{X}^+\boldsymbol{\Delta}_t^+)}{2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)} \\ &\quad = \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^+\boldsymbol{\Delta}_t^+\boldsymbol{X}^+\boldsymbol{\Delta}_t^+) + c_\star\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}\boldsymbol{A}_{\mathsf{P}}) - k_{b_\star}\,\|[\boldsymbol{A}_{\mathsf{P}}, \boldsymbol{X}]\|_F^2}{2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)} \\ &\quad \le r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star).\end{aligned}$$

The denominator is positive because $\boldsymbol{A}_{\mathsf{P}} \succ 0$ and $\boldsymbol{X} \neq 0$. Recall that δ_{cov} is the largest normalized change in the noise term over these full-space test matrices:

$$\delta_{\mathrm{cov}} = \max_{\substack{\boldsymbol{X} \in \mathbb{S}_d^+ \\ \|\boldsymbol{X}\|_F = 1}} \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) - \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^+\boldsymbol{\Delta}_t^+\boldsymbol{X}^+\boldsymbol{\Delta}_t^+)}{2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)}.$$

Adding back the subtracted fraction and bounding it by this maximum, as defined in (53), gives

$$\begin{aligned}q_{\boldsymbol{A}_{\mathsf{P}}, b_\star}(\boldsymbol{X}) &\le r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star) + \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) - \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^+\boldsymbol{\Delta}_t^+\boldsymbol{X}^+\boldsymbol{\Delta}_t^+)}{2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)} \\ &\le r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star) + \delta_{\mathrm{cov}}.\end{aligned}$$

This holds for every full-space test matrix. Maximizing both sides,

$$\begin{aligned}r_{\boldsymbol{A}_{\mathsf{P}}}(b_\star) &= \max_{\substack{\boldsymbol{X} \in \mathbb{S}_d^+ \\ \|\boldsymbol{X}\|_F = 1}} q_{\boldsymbol{A}_{\mathsf{P}}, b_\star}(\boldsymbol{X}) \\ &\le \max_{\substack{\boldsymbol{X} \in \mathbb{S}_d^+ \\ \|\boldsymbol{X}\|_F = 1}} \big[r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star) + \delta_{\mathrm{cov}} \big] \\ &= r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star) + \delta_{\mathrm{cov}}.\end{aligned} \tag{80}$$

We can now combine the lower bound from the embedded test matrices and the upper bound (80) with Step 1:

$$\begin{aligned}r_{\boldsymbol{A}_{\mathsf{P}}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) &\ge r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) \\ &\ge \Phi(\mu_\star), \\ r_{\boldsymbol{A}_{\mathsf{P}}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) &\le r_{\boldsymbol{A}^+ + \lambda_{\mathsf{P}} \boldsymbol{I}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) + \delta_{\mathrm{cov}} \\ &\le \Phi(L) + \delta_{\mathrm{cov}}.\end{aligned}$$

Equivalently,

$$\Phi(\mu_\star) \le r_{\boldsymbol{A}_\mathsf{P}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) \le \Phi(L) + \delta_{\mathrm{cov}}. \tag{81}$$

Step 3. Compare the critical LRs.

Recall that $b_\star^\mathsf{P} = \beta_1(1 - \eta_\star^\mathsf{P}\lambda_{\mathrm{WD}})$ and $b_\star = \beta_1(1 - \eta_\star\lambda_{\mathrm{WD}})$ are the momenta at the two critical LRs. The critical-rate identity (72) expresses each quotient maximum as the inverse critical rate plus the same weight-decay offset:

$$r_{\boldsymbol{A}_\mathsf{P}}(b_\star^\mathsf{P}) = \frac{1}{\eta_\star^\mathsf{P}} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1},$$
$$r_{\boldsymbol{A}^+}(b_\star) = \frac{1}{\eta_\star} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}.$$

Subtracting the same offset from both thresholds,

$$\begin{aligned}\frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star} &= \left[r_{\boldsymbol{A}_\mathsf{P}}(b_\star^\mathsf{P}) - \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}\right] - \left[r_{\boldsymbol{A}^+}(b_\star) - \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}\right] \\ &= r_{\boldsymbol{A}_\mathsf{P}}(b_\star^\mathsf{P}) - r_{\boldsymbol{A}^+}(b_\star).\end{aligned}$$

The two critical momenta differ by

$$\begin{aligned}b_\star^\mathsf{P} - b_\star &= \beta_1(1 - \eta_\star^\mathsf{P}\lambda_{\mathrm{WD}}) - \beta_1(1 - \eta_\star\lambda_{\mathrm{WD}}) \\ &= \beta_1\lambda_{\mathrm{WD}}(\eta_\star - \eta_\star^\mathsf{P}).\end{aligned}$$

Also, both critical LRs are positive, so

$$\frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star} = \frac{\eta_\star - \eta_\star^\mathsf{P}}{\eta_\star^\mathsf{P}\eta_\star}.$$

These two identities determine the signs in the following cases.

If $\eta_\star^\mathsf{P} \le \eta_\star$, then $b_\star^\mathsf{P} \ge b_\star$ and $1/\eta_\star^\mathsf{P} - 1/\eta_\star \ge 0$. Since $r_{\boldsymbol{A}_\mathsf{P}}$ is nonincreasing in its momentum argument, $r_{\boldsymbol{A}_\mathsf{P}}(b_\star^\mathsf{P}) \le r_{\boldsymbol{A}_\mathsf{P}}(b_\star)$. Combining this with the upper bound in (81),

$$\begin{aligned}\frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star} &= r_{\boldsymbol{A}_\mathsf{P}}(b_\star^\mathsf{P}) - r_{\boldsymbol{A}^+}(b_\star) \\ &\le r_{\boldsymbol{A}_\mathsf{P}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) \\ &\le \Phi(L) + \delta_{\mathrm{cov}}.\end{aligned}$$

In this case both clipped bounds follow from

$$\min\{\Phi(\mu_\star), 0\} \le 0 \le \frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star},$$
$$\frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star} \le \Phi(L) + \delta_{\mathrm{cov}} \le \max\{\Phi(L) + \delta_{\mathrm{cov}}, 0\}.$$

If $\eta_\star^\mathsf{P} \ge \eta_\star$, then $b_\star^\mathsf{P} \le b_\star$ and $1/\eta_\star^\mathsf{P} - 1/\eta_\star \le 0$. Monotonicity now gives $r_{\boldsymbol{A}_\mathsf{P}}(b_\star^\mathsf{P}) \ge r_{\boldsymbol{A}_\mathsf{P}}(b_\star)$. Combining this with the lower bound in (81),

$$\begin{aligned}\frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star} &= r_{\boldsymbol{A}_\mathsf{P}}(b_\star^\mathsf{P}) - r_{\boldsymbol{A}^+}(b_\star) \\ &\ge r_{\boldsymbol{A}_\mathsf{P}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star) \\ &\ge \Phi(\mu_\star).\end{aligned}$$

The clipped bounds now follow from

$$\min\{\Phi(\mu_\star), 0\} \le \Phi(\mu_\star) \le \frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star},$$
$$\frac{1}{\eta_\star^\mathsf{P}} - \frac{1}{\eta_\star} \le 0 \le \max\{\Phi(L) + \delta_{\mathrm{cov}}, 0\}.$$

This proves (76) in both cases.

Finally, if $\beta_1\lambda_{\mathrm{WD}} = 0$, then

$$b_\star^{\mathsf{P}} = \beta_1 - \beta_1\lambda_{\mathrm{WD}}\eta_\star^{\mathsf{P}} = \beta_1 = \beta_1 - \beta_1\lambda_{\mathrm{WD}}\eta_\star = b_\star.$$

The critical identities and (81) therefore give

$$\begin{aligned}\Phi(\mu_\star) &\le r_{\boldsymbol{A}_{\mathsf{P}}}(b_\star) - r_{\boldsymbol{A}^+}(b_\star)\\ &= r_{\boldsymbol{A}_{\mathsf{P}}}(b_\star^{\mathsf{P}}) - r_{\boldsymbol{A}^+}(b_\star)\\ &= \frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star}\\ &\le \Phi(L) + \delta_{\mathrm{cov}}.\end{aligned}$$

Hence no clipping is needed in this case. This completes the proof of Proposition G.5. □

A scalar example. Let $\beta_1 = \lambda_{\mathrm{WD}} = 0$, $d = p = 1$, $\boldsymbol{A} = 1$, $\mathbb{E}\boldsymbol{\Delta}_t^2 = 1$, and $\lambda_{\mathsf{P}} = 0.01$. For the homogeneous scalar SGD recurrence with mean $a > 0$ and centered noise of variance σ^2, the factor by which one step multiplies the second moment is

$$\begin{aligned}\mathbb{E}\big(1 - \eta(a + \Delta_t)\big)^2 &= 1 - 2\eta a + \eta^2\mathbb{E}(a + \Delta_t)^2\\ &= 1 - 2\eta a + \eta^2(a^2 + \sigma^2).\end{aligned}$$

For $\eta > 0$, this factor is below one if and only if

$$-2\eta a + \eta^2(a^2 + \sigma^2) < 0 \quad \Longleftrightarrow \quad \eta < \frac{2a}{a^2 + \sigma^2}.$$

The inverse critical rate is therefore $(a^2 + \sigma^2)/(2a)$. In the projected system $a = 1$ and $\sigma^2 = 1$, while compensation replaces a by $1 + \lambda_{\mathsf{P}}$ and leaves the variance unchanged. Hence

$$\begin{aligned}\frac{1}{\eta_\star} &= 1,\\ \frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star} &= \frac{(1 + \lambda_{\mathsf{P}})^2 + 1}{2(1 + \lambda_{\mathsf{P}})} - 1\\ &= \frac{\lambda_{\mathsf{P}}^2}{2(1 + \lambda_{\mathsf{P}})} \approx 4.95 \cdot 10^{-5}.\end{aligned}$$

Here $\mu_\star = L = 1$ and $\delta_{\mathrm{cov}} = 0$. Since $b_\star = 0$, $c_\star = 1$ and $r_{\boldsymbol{A}^+}(b_\star) = 1$, the definition of Φ gives

$$\Phi(1) = \frac{\lambda_{\mathsf{P}}}{2} - \frac{\lambda_{\mathsf{P}}}{1 + \lambda_{\mathsf{P}}}\left(1 - \frac{1}{2}\right) = \frac{\lambda_{\mathsf{P}}^2}{2(1 + \lambda_{\mathsf{P}})}.$$

Both compensation endpoints therefore equal the inverse-rate difference. The comparison is exact because the shared noise is compared directly rather than bounded separately in each system.

The SGD case. For SGD without weight decay, $\beta_1 = \lambda_{\mathrm{WD}} = 0$ and $\boldsymbol{P} = \boldsymbol{I}$, so $\boldsymbol{A} = \boldsymbol{H}_{t_0}$. Here $b_\star^{\mathsf{P}} = 0$, $\alpha_\star^{\mathsf{P}} = \eta_\star^{\mathsf{P}}$ and $c_\star = 1$. The critical identity (72) gives $r_{\boldsymbol{A}^+}(0) = 1/\eta_\star$. Substituting these values into the definitions of ρ_T, H_T and Φ gives

$$\begin{aligned}\rho_T &= M^{2/T}, \qquad H_T = \frac{M^{2/T} - 1}{2\mu_\star^{\mathsf{P}}(\eta_\star^{\mathsf{P}})^2},\\ \Phi(s) &= \frac{\lambda_{\mathsf{P}}}{2} - \frac{\lambda_{\mathsf{P}}}{s + \lambda_{\mathsf{P}}}\left(\frac{1}{\eta_\star} - \frac{s}{2}\right).\end{aligned}$$

The theorem applies without clipping. The direct LR bound derived in Appendix G.4 is $\widehat{\eta}_\star(M, T) \le \eta_\star^{\mathsf{P}} + (M^{2/T} - 1)/(2\mu_\star^{\mathsf{P}})$.

The covariance mismatch can be bounded by second moments of the discarded and cross blocks of the noise. Recall that $\boldsymbol{\Delta}_t^{\perp} = \boldsymbol{\Delta}_t - \Pi^+\boldsymbol{\Delta}_t\Pi^+$ contains precisely those discarded and cross blocks.

Lemma G.7 (A bound on the covariance mismatch)**.** *Under Assumption D.14, let* $v_0, v_E \geq 0$ *satisfy* $\mathbb{E}(\boldsymbol{\Delta}_t^+)^2 \preceq v_0(\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I})$ *and* $\mathbb{E}(\boldsymbol{\Delta}_t^\perp)^2 \preceq v_E\boldsymbol{A}_{\mathsf{P}}$*. Then*

$$\delta_{\mathrm{cov}} \leq \sqrt{v_0 v_E} + \frac{v_E}{2}.$$

Proof of Lemma G.7. We bound the noise difference that defines δ_{cov}, keeping its denominator $2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)$ unchanged. Within this proof write

$$\boldsymbol{\Delta}_t^\circ := \Pi^+\boldsymbol{\Delta}_t\Pi^+ = \mathrm{diag}(\boldsymbol{\Delta}_t^+, 0), \qquad \boldsymbol{\Delta}_t = \boldsymbol{\Delta}_t^\circ + \boldsymbol{\Delta}_t^\perp.$$

Step 1. Bound the separate noise terms.

The retained-noise moment bound in the hypothesis extends to the full space:

$$\begin{aligned}\mathbb{E}(\boldsymbol{\Delta}_t^\circ)^2 &= \mathrm{diag}\big(\mathbb{E}(\boldsymbol{\Delta}_t^+)^2, 0\big)\\ &\preceq v_0\,\mathrm{diag}(\boldsymbol{A}^+ + \lambda_{\mathsf{P}}\boldsymbol{I}, 0)\\ &\preceq v_0\boldsymbol{A}_{\mathsf{P}}.\end{aligned}$$

The last inequality uses the positive definiteness of the discarded block when it is present, and is an equality when $p = d$. Fix $0 \neq \boldsymbol{X} \in \mathbb{S}_d^+$. Since $\boldsymbol{\Delta}_t^\circ = \mathrm{diag}(\boldsymbol{\Delta}_t^+, 0)$, block multiplication gives

$$\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^+\boldsymbol{\Delta}_t^+\boldsymbol{X}^+\boldsymbol{\Delta}_t^+) = \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t^\circ\boldsymbol{X}\boldsymbol{\Delta}_t^\circ).$$

We first estimate the cross term and the discarded-noise term separately. For any symmetric $\boldsymbol{\Delta}$, symmetry and cyclicity of the trace give

$$\begin{aligned}\|\boldsymbol{\Delta}\boldsymbol{X}\|_F^2 &= \mathrm{tr}\big((\boldsymbol{\Delta}\boldsymbol{X})^\top\boldsymbol{\Delta}\boldsymbol{X}\big) && \text{Frobenius norm}\\ &= \mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}^2\boldsymbol{X}) && \text{symmetry}\\ &= \mathrm{tr}(\boldsymbol{X}^2\boldsymbol{\Delta}^2), && \text{cyclicity of the trace}\\ \|\boldsymbol{X}\boldsymbol{\Delta}\|_F^2 &= \mathrm{tr}\big((\boldsymbol{X}\boldsymbol{\Delta})^\top\boldsymbol{X}\boldsymbol{\Delta}\big) && \text{Frobenius norm}\\ &= \mathrm{tr}(\boldsymbol{\Delta}\boldsymbol{X}^2\boldsymbol{\Delta}) && \text{symmetry}\\ &= \mathrm{tr}(\boldsymbol{X}^2\boldsymbol{\Delta}^2). && \text{cyclicity of the trace}\end{aligned}$$

The two moment bounds therefore imply

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{\Delta}_t^\circ\boldsymbol{X}\|_F^2 &= \mathrm{tr}\big(\boldsymbol{X}^2\mathbb{E}(\boldsymbol{\Delta}_t^\circ)^2\big) && \text{the norm identity above}\\ &\leq v_0\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2), && \text{the retained-noise moment bound}\\ \mathbb{E}\,\big\|\boldsymbol{X}\boldsymbol{\Delta}_t^\perp\big\|_F^2 &= \mathrm{tr}\big(\boldsymbol{X}^2\mathbb{E}(\boldsymbol{\Delta}_t^\perp)^2\big) && \text{the norm identity above}\\ &\leq v_E\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2). && \text{the discarded-noise moment bound}\end{aligned}$$

Both inequalities use the same trace-order fact. For $\boldsymbol{Q} \succeq 0$,

$$\mathrm{tr}(\boldsymbol{X}^2\boldsymbol{Q}) = \mathrm{tr}(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{X}) \geq 0.$$

Apply this to the differences between the right and left matrices in the two moment bounds.

For the cross term, the Cauchy–Schwarz inequality in the Frobenius inner product and then in L^2 gives

$$\begin{aligned}&\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t^\circ\boldsymbol{X}\boldsymbol{\Delta}_t^\perp)\\ &\quad= \mathbb{E}\langle\boldsymbol{\Delta}_t^\circ\boldsymbol{X}, \boldsymbol{X}\boldsymbol{\Delta}_t^\perp\rangle_F && \text{symmetry of } \boldsymbol{X} \text{ and } \boldsymbol{\Delta}_t^\circ\\ &\quad\leq \mathbb{E}\big[\|\boldsymbol{\Delta}_t^\circ\boldsymbol{X}\|_F\,\big\|\boldsymbol{X}\boldsymbol{\Delta}_t^\perp\big\|_F\big] && \text{Frobenius Cauchy–Schwarz}\\ &\quad\leq \sqrt{\mathbb{E}\,\|\boldsymbol{\Delta}_t^\circ\boldsymbol{X}\|_F^2\;\mathbb{E}\,\big\|\boldsymbol{X}\boldsymbol{\Delta}_t^\perp\big\|_F^2} && \text{Cauchy–Schwarz in } L^2\\ &\quad\leq \sqrt{v_0 v_E}\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2). && \text{the two moment bounds}\end{aligned}$$

For the discarded-noise term, the two Frobenius norms are equal, so

$$\begin{aligned}
&\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}\boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}) \\
&\quad= \mathbb{E}\langle \boldsymbol{\Delta}_t^{\perp}\boldsymbol{X}, \boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}\rangle_F \\
&\quad\le \mathbb{E}\big[\big\|\boldsymbol{\Delta}_t^{\perp}\boldsymbol{X}\big\|_F \big\|\boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}\big\|_F\big] && \text{Frobenius Cauchy–Schwarz} \\
&\quad= \mathbb{E}\,\big\|\boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}\big\|_F^2 && \text{the norm identity above} \\
&\quad\le v_E\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2). && \text{the discarded-noise moment bound}
\end{aligned}$$

Step 2. Combine the terms and normalize.

Expand $\boldsymbol{\Delta}_t = \boldsymbol{\Delta}_t^{\circ} + \boldsymbol{\Delta}_t^{\perp}$ in the numerator. The two cross terms have the same trace by cyclicity, and the preceding bounds give

$$\begin{aligned}
&\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) - \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t^{\circ}\boldsymbol{X}\boldsymbol{\Delta}_t^{\circ}) \\
&\quad= 2\,\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t^{\circ}\boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}) + \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}\boldsymbol{X}\boldsymbol{\Delta}_t^{\perp}) \\
&\quad\le \big(2\sqrt{v_0 v_E} + v_E\big)\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2).
\end{aligned}$$

For every nonzero PSD $\boldsymbol{X}$, the denominator $2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)$ is positive. Dividing by it and using the retained-block trace identity above,

$$\frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) - \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^{+}\boldsymbol{\Delta}_t^{+}\boldsymbol{X}^{+}\boldsymbol{\Delta}_t^{+})}{2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)} \le \frac{2\sqrt{v_0 v_E} + v_E}{2} = \sqrt{v_0 v_E} + \frac{v_E}{2}.$$

The right side is independent of $\boldsymbol{X}$, so it also bounds the maximum δ_{cov}. This completes the proof of Lemma G.7. □

Since $\boldsymbol{A}^{+} \preceq \boldsymbol{A}^{+} + \lambda_{\mathsf{P}}\boldsymbol{I}$, one may take v_0 to be the noise level of the projected system, the smallest v with $\mathbb{E}(\boldsymbol{\Delta}_t^{+})^2 \preceq v\boldsymbol{A}^{+}$. The lemma controls the noise in the discarded and cross blocks relative to the compensated effective curvature. A bound on the compensation coefficient alone does not control these quantities, as the discarded-noise example below shows.

Dependence on saddleness. The next bound makes the dependence on the compensation coefficient explicit. It bounds the Hessian noise in each system by its second moment relative to the corresponding positive definite mean.

Corollary G.8 (Compensation and saddleness)**.** *Under Assumption D.14, let $v \ge 0$ satisfy*

$$\mathbb{E}(\boldsymbol{\Delta}_t^{+})^2 \preceq v\boldsymbol{A}^{+}, \qquad \mathbb{E}\boldsymbol{\Delta}_t^{2} \preceq v\boldsymbol{A}_{\mathsf{P}}. \tag{82}$$

Then the projected and compensated critical LRs satisfy

$$\left|\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star}\right| \le \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}} + \frac{v}{2}. \tag{83}$$

If Assumption D.15 also holds, with the search window specified in Appendix D.4, then the population detector satisfies

$$\left|\frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star}\right| \le \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}} + \frac{v}{2} + H_T, \tag{84}$$

where H_T is the finite-horizon term in (56).

A finite v always exists because both means are positive definite and the Hessian noise has finite second moments. The corollary does not assume that v is small. Its second condition uses $\boldsymbol{A}_{\mathsf{P}}$, so v can depend on the compensation coefficient and need not stay bounded as that coefficient decreases.

Recall that the effective saddleness in (15) is

$$\nu = \frac{\max\{0, -\lambda_{\min}(\boldsymbol{A})\}}{L}, \qquad L = \lambda_{\max}(\boldsymbol{A}) > 0.$$

When this saddleness is positive, choosing $\lambda_{\mathsf{P}} = c\nu L$ with $c > 1$ makes $\boldsymbol{A}_{\mathsf{P}}$ positive definite. More generally, whenever a valid compensation satisfies $\lambda_{\mathsf{P}} \le c\nu L$, (83) gives

$$\left|\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star}\right| \le \frac{c(1-\beta_1)}{2(1+\beta_1)}\nu L + \frac{v}{2}.$$

Thus saddleness controls the curvature shift, while v accounts for Hessian noise. If the compensation is chosen from an estimated eigenvalue, an upper bound on λ_{P} must also retain the estimation error. Small saddleness alone does not control v.

Proof of Corollary G.8. We first bound the change in a critical LR caused by adding noise to a positive definite mean. Applying this bound to the projected and compensated means leaves an explicit curvature shift and at most $v/2$ from the difference of the two noise contributions.

Step 1. Bound the noise contribution to the quotient.

Consider the AGM recurrence with a generic mean $\boldsymbol{C} \succ 0$ and symmetric centered noise satisfying $\mathbb{E}\boldsymbol{\Delta}_t^2 \preceq v\boldsymbol{C}$. Recall that $L_{\boldsymbol{C}} = \lambda_{\max}(\boldsymbol{C})$, $c_b = (1-b)/(1+b)$, and $r_{\boldsymbol{C}}(b)$ is the maximum of the stability quotient over nonzero positive semidefinite test matrices. For such a matrix $\boldsymbol{X}$, the noise numerator is nonnegative because

$$\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) = \mathbb{E}\left\|\boldsymbol{X}^{1/2}\boldsymbol{\Delta}_t\boldsymbol{X}^{1/2}\right\|_F^2 \geq 0.$$

For the upper bound, apply the Cauchy–Schwarz inequality in the Frobenius inner product and use the symmetry of $\boldsymbol{X}$ and $\boldsymbol{\Delta}_t$:

$$\begin{aligned}
\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) &= \mathbb{E}\langle\boldsymbol{\Delta}_t\boldsymbol{X}, \boldsymbol{X}\boldsymbol{\Delta}_t\rangle_F && \text{the Frobenius inner product}\\
&\leq \mathbb{E}\left[\|\boldsymbol{\Delta}_t\boldsymbol{X}\|_F \|\boldsymbol{X}\boldsymbol{\Delta}_t\|_F\right] && \text{Cauchy–Schwarz}\\
&= \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}^2\boldsymbol{\Delta}_t^2) && \text{the two norms are equal}\\
&= \mathrm{tr}\left(\boldsymbol{X}^2\mathbb{E}\boldsymbol{\Delta}_t^2\right) && \text{linearity of expectation}\\
&\leq v\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2). && \mathbb{E}\boldsymbol{\Delta}_t^2 \preceq v\boldsymbol{C}
\end{aligned}$$

The quotient divides this numerator by $2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2) > 0$. Its noise contribution therefore lies between zero and $v/2$ for every admissible test matrix. The maximum without noise equals $c_b L_{\boldsymbol{C}}/2$ by (73). Taking maxima gives, for every $-1 < b < 1$,

$$\frac{c_b L_{\boldsymbol{C}}}{2} \leq r_{\boldsymbol{C}}(b) \leq \frac{c_b L_{\boldsymbol{C}}}{2} + \frac{v}{2}.$$

Step 2. Compare with the critical LR without noise.

Let η_{crit} be the critical rate for this noisy recurrence. Recall from (74) that the same mean without noise has critical rate

$$\eta_{\mathrm{det}} = \frac{2(1+\beta_1)}{(1-\beta_1)L_{\boldsymbol{C}} + 2\beta_1\lambda_{\mathrm{WD}}}.$$

By the exact stability criterion (Theorem G.1), adding the nonnegative noise term can only remove strictly stable LRs of the homogeneous recurrence. Hence $\eta_{\mathrm{crit}} \leq \eta_{\mathrm{det}}$. Their effective momenta are

$$b_{\mathrm{crit}} \coloneqq \beta_1(1-\eta_{\mathrm{crit}}\lambda_{\mathrm{WD}}), \qquad b_{\mathrm{det}} \coloneqq \beta_1(1-\eta_{\mathrm{det}}\lambda_{\mathrm{WD}}).$$

Both lie in $(-1, 1)$ by Corollary G.4, and

$$\begin{aligned}
b_{\mathrm{crit}} - b_{\mathrm{det}} &= \beta_1\lambda_{\mathrm{WD}}(\eta_{\mathrm{det}} - \eta_{\mathrm{crit}})\\
&\geq 0.
\end{aligned}$$

Recall that the critical-rate identity (72) adds the same weight-decay offset to the inverse rate in both systems. Since $r_{\boldsymbol{C}}$ is nonincreasing in its momentum argument,

$$\begin{aligned}
\frac{1}{\eta_{\mathrm{crit}}} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1} &= r_{\boldsymbol{C}}(b_{\mathrm{crit}}) && \text{critical-rate identity}\\
&\leq r_{\boldsymbol{C}}(b_{\mathrm{det}}) && \text{monotonicity in } b\\
&\leq \frac{c_{b_{\mathrm{det}}} L_{\boldsymbol{C}}}{2} + \frac{v}{2} && \text{Step 1}\\
&= \frac{1}{\eta_{\mathrm{det}}} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1} + \frac{v}{2}. && \text{boundary without noise}
\end{aligned}$$

Subtracting the common offset and using $\eta_{\mathrm{crit}} \le \eta_{\mathrm{det}}$ gives

$$0 \le \frac{1}{\eta_{\mathrm{crit}}} - \frac{1}{\eta_{\mathrm{det}}} \le \frac{v}{2}.$$

This argument uses monotonicity on the full interval $(-1, 1)$ and therefore also covers negative effective critical momenta.

Step 3. Apply the bound to the two systems.

The projected system has mean $\boldsymbol{A}^+$, noise $\boldsymbol{\Delta}_t^+$ and largest mean eigenvalue L. The compensated system has mean $\boldsymbol{A}_{\mathsf{P}}$, noise $\boldsymbol{\Delta}_t$ and largest mean eigenvalue $L + \lambda_{\mathsf{P}}$. The two inequalities in (82) allow us to apply Step 2 with the same v:

$$0 \le \frac{1}{\eta_\star} - \frac{(1-\beta_1)L + 2\beta_1\lambda_{\mathrm{WD}}}{2(1+\beta_1)} \le \frac{v}{2},$$
$$0 \le \frac{1}{\eta_\star^{\mathsf{P}}} - \frac{(1-\beta_1)(L+\lambda_{\mathsf{P}}) + 2\beta_1\lambda_{\mathrm{WD}}}{2(1+\beta_1)} \le \frac{v}{2}.$$

The difference between the two displayed noise contributions lies in $[-v/2, v/2]$. The deterministic terms differ by

$$\begin{aligned}
&\frac{(1-\beta_1)(L+\lambda_{\mathsf{P}}) + 2\beta_1\lambda_{\mathrm{WD}}}{2(1+\beta_1)} - \frac{(1-\beta_1)L + 2\beta_1\lambda_{\mathrm{WD}}}{2(1+\beta_1)} \\
&\quad = \frac{(1-\beta_1)(L+\lambda_{\mathsf{P}}) - (1-\beta_1)L}{2(1+\beta_1)} && \text{cancel the weight-decay terms} \\
&\quad = \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}}.
\end{aligned}$$

To subtract the two noise contributions, use the upper endpoint of one interval and the lower endpoint of the other:

$$\begin{aligned}
&\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star} - \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}} \\
&\quad = \left[\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{(1-\beta_1)(L+\lambda_{\mathsf{P}}) + 2\beta_1\lambda_{\mathrm{WD}}}{2(1+\beta_1)}\right] \\
&\qquad - \left[\frac{1}{\eta_\star} - \frac{(1-\beta_1)L + 2\beta_1\lambda_{\mathrm{WD}}}{2(1+\beta_1)}\right] \\
&\quad \le \frac{v}{2} - 0 = \frac{v}{2}. && \text{each bracket lies in } [0, v/2]
\end{aligned}$$

Using the opposite endpoints in the same identity gives

$$\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star} - \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}} \ge 0 - \frac{v}{2} = -\frac{v}{2}.$$

Since the curvature shift is nonnegative, the triangle inequality gives

$$\begin{aligned}
\left|\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star}\right| &\le \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}} + \left|\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta_\star} - \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}}\right| \\
&\le \frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}} + \frac{v}{2}.
\end{aligned}$$

This proves (83).

For the population detector, Proposition G.6 bounds the inverse of every accepted rate below by $1/\eta_\star^{\mathsf{P}} - H_T$. Taking their infimum gives the same lower bound for $1/\widehat{\eta}_\star(M,T)$. Acceptance of $\eta_\star^{\mathsf{P}}$ in Assumption D.15 gives $\widehat{\eta}_\star(M,T) \ge \eta_\star^{\mathsf{P}}$. Together these imply

$$-H_T \le \frac{1}{\widehat{\eta}_\star(M,T)} - \frac{1}{\eta_\star^{\mathsf{P}}} \le 0.$$

Split the inverse-rate error at $\eta_\star^{\mathsf{P}}$ and apply the triangle inequality:

$$\begin{aligned}\left|\frac{1}{\widehat{\eta}_\star(M,T)}-\frac{1}{\eta_\star}\right| &= \left|\left(\frac{1}{\widehat{\eta}_\star(M,T)}-\frac{1}{\eta_\star^{\mathsf{P}}}\right)+\left(\frac{1}{\eta_\star^{\mathsf{P}}}-\frac{1}{\eta_\star}\right)\right| \\ &\le \left|\frac{1}{\widehat{\eta}_\star(M,T)}-\frac{1}{\eta_\star^{\mathsf{P}}}\right|+\left|\frac{1}{\eta_\star^{\mathsf{P}}}-\frac{1}{\eta_\star}\right| \\ &\le H_T+\frac{1-\beta_1}{2(1+\beta_1)}\lambda_{\mathsf{P}}+\frac{v}{2}.\end{aligned}$$

This proves (84) and completes the proof of Corollary G.8. □

Small saddleness alone is insufficient. For $0<s\le 1$, take $\boldsymbol{P}=\boldsymbol{I}_2$, $\beta_1=\lambda_{\mathrm{WD}}=0$,

$$\boldsymbol{H}_{\mathcal{B}_t}=\mathrm{diag}(1,-s+\sqrt{2s}\,\xi_t),\qquad \Pr(\xi_t=\pm 1)=\tfrac{1}{2},\qquad \lambda_{\mathsf{P}}=2s.$$

The saddleness is s. The retained coordinate is noiseless, while the compensated second coordinate has mean s and noise variance $2s$. Here $\boldsymbol{A}_{\mathsf{P}}=\mathrm{diag}(1+2s,s)$ and $\boldsymbol{\Delta}_t^{+}=0$. To evaluate the covariance mismatch, recall that δ_{cov} maximizes the normalized noise difference in (53). For a symmetric 2×2 test matrix with entries X_{ij},

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) &= 2sX_{22}^2,\\ \mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2) &= (1+2s)(X_{11}^2+X_{12}^2)+s(X_{12}^2+X_{22}^2),\\ \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t)}{2\,\mathrm{tr}(\boldsymbol{A}_{\mathsf{P}}\boldsymbol{X}^2)} &= \frac{sX_{22}^2}{(1+2s)(X_{11}^2+X_{12}^2)+s(X_{12}^2+X_{22}^2)}\le 1.\end{aligned}$$

Equality holds at the admissible matrix $\boldsymbol{X}=\mathrm{diag}(0,1)$, so $\delta_{\mathrm{cov}}=1$. The scalar threshold above and the definition of Φ now give

$$\begin{aligned}\frac{1}{\eta_\star} &= \frac{1}{2},\qquad \Phi(1)=s,\qquad \delta_{\mathrm{cov}}=1,\\ \frac{1}{\eta_\star^{\mathsf{P}}} &= \max\left\{\frac{1+2s}{2},1+\frac{s}{2}\right\}=1+\frac{s}{2}.\end{aligned}$$

The last equality uses $s\le 1$. The inverse-rate difference tends to $1/2$ as $s\to 0$. The discarded noise prevents the compensation error from vanishing, despite vanishing saddleness.

G.4 Finite-Horizon Error

The homogeneous compensated recurrence, with mean $\boldsymbol{A}_{\mathsf{P}}=\boldsymbol{A}+\lambda_{\mathsf{P}}\boldsymbol{I}$ and centered noise $\boldsymbol{\Delta}_t=\boldsymbol{A}_{\mathcal{B}_t}-\boldsymbol{A}$, is

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \big[(1+b)\boldsymbol{I}-\alpha(\boldsymbol{A}_{\mathsf{P}}+\boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t-b\boldsymbol{x}_{t-1},\\ b &= \beta_1(1-\eta\lambda_{\mathrm{WD}}),\qquad \alpha=\eta(1-\beta_1),\end{aligned}$$

Its second-moment map on the history state $\boldsymbol{y}_t=(\boldsymbol{x}_t,\boldsymbol{x}_{t-1})^\top$ is $\mathcal{T}_\eta^{\mathsf{P}}$. The mean and noise law stay fixed as the candidate LR η varies.

Acceptance means $\mathbb{E}\,\|\boldsymbol{\epsilon}_t\|^2\le M^2$ for every unit start and $0\le t\le T$ in the original affine run. We convert this condition into a moment-radius bound, then apply Proposition G.2. Recall the uniform radius bound from (55),

$$\rho_T=\left[\frac{M(1+\beta_1)}{1-\beta_1}\right]^{2/T}.$$

It depends on M and T, not on the candidate LR. The last proposition gives a sufficient cutoff for the original affine run to be accepted at $\eta_\star^{\mathsf{P}}$.

G.4.1 Bounding the Moment Radius

Lemma G.9 (The expected-displacement criterion bounds the moment radius)**.** *Under Assumption D.14, every LR $\eta\in\mathcal{I}$ accepted by the detector (48) satisfies*

$$\rho(\mathcal{T}_\eta^{\mathsf{P}})\le\left[\frac{M(1+\beta_1|1-\eta\lambda_{\mathrm{WD}}|)}{1-\beta_1}\right]^{2/T}\le\rho_T. \tag{85}$$

Proof of Lemma G.9. Fix an accepted LR η and write $\rho := \rho(\mathcal{T}_\eta^{\mathsf{P}})$. If $\rho < 1$, the claim follows because the proposed upper bound is greater than one. We therefore assume $\rho \geq 1$. Our goal is to prove

$$\rho^{T/2} \leq \frac{M(1+b)}{1-\beta_1}, \qquad b = \beta_1(1-\eta\lambda_{\mathrm{WD}}).$$

The detector bounds displacement, whereas ρ describes growth of the full second moment. We connect the two by choosing a nonnegative quadratic form whose expectation grows by the factor ρ at each step. The three steps below remove the affine forcing, construct this quadratic form, and bound its value at time T using the detector.

Step 1. Bound the homogeneous trajectory.

For any unit vector $\boldsymbol{v}$, run the original affine simulation (47) twice, from initial displacements $\boldsymbol{v}$ and $-\boldsymbol{v}$, using the same mini-batches and the same checkpoint momentum. Denote their positions and momenta by $(\boldsymbol{\epsilon}_t^+, \boldsymbol{m}_t^+)$ and $(\boldsymbol{\epsilon}_t^-, \boldsymbol{m}_t^-)$. Within this proof, set

$$\boldsymbol{\epsilon}_t := \frac{\boldsymbol{\epsilon}_t^+ - \boldsymbol{\epsilon}_t^-}{2}, \qquad \boldsymbol{p}_t := \frac{\eta}{2}\boldsymbol{P}^{-1}(\boldsymbol{m}_t^+ - \boldsymbol{m}_t^-).$$

Here $\boldsymbol{p}_t$ is the momentum difference scaled into displacement units. Thus $\boldsymbol{\epsilon}_0 = \boldsymbol{v}$ and $\boldsymbol{p}_0 = 0$. Subtracting the two momentum updates cancels their common batch gradient. Subtracting the two position updates also cancels the checkpoint weight-decay drift. The resulting equations are

$$\begin{aligned}\boldsymbol{p}_{t+1} &= \beta_1\boldsymbol{p}_t + \alpha(\boldsymbol{P}^{-1}\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I})\boldsymbol{\epsilon}_t,\\ \boldsymbol{\epsilon}_{t+1} &= (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{\epsilon}_t - \boldsymbol{p}_{t+1}, \qquad \alpha = \eta(1-\beta_1).\end{aligned}$$

Here $\boldsymbol{H}_{\mathcal{B}_t}$ is the raw mini-batch Hessian. Weight decay remains in the separate position factor $1-\eta\lambda_{\mathrm{WD}}$. These are the same homogeneous compensated dynamics as before, expressed in the detector's original coordinates.

Both affine runs pass the detector because their initial displacements have unit norm. Consequently, for every $0 \leq t \leq T$,

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{\epsilon}_t\|^2 &= \frac{1}{4}\mathbb{E}\,\|\boldsymbol{\epsilon}_t^+ - \boldsymbol{\epsilon}_t^-\|^2\\ &\leq \frac{1}{2}\mathbb{E}\big(\|\boldsymbol{\epsilon}_t^+\|^2 + \|\boldsymbol{\epsilon}_t^-\|^2\big)\\ &\leq \frac{1}{2}(M^2 + M^2) = M^2.\end{aligned}$$

The second line uses $\|\boldsymbol{a}-\boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$. To control the momentum at time T, use the preceding position update:

$$\boldsymbol{p}_T = (1-\eta\lambda_{\mathrm{WD}})\boldsymbol{\epsilon}_{T-1} - \boldsymbol{\epsilon}_T.$$

The search window guarantees $0 \leq 1-\eta\lambda_{\mathrm{WD}} \leq 1$, and $T \geq 1$. Minkowski's inequality therefore gives

$$\begin{aligned}(\mathbb{E}\,\|\boldsymbol{p}_T\|^2)^{1/2} &\leq (1-\eta\lambda_{\mathrm{WD}})(\mathbb{E}\,\|\boldsymbol{\epsilon}_{T-1}\|^2)^{1/2} + (\mathbb{E}\,\|\boldsymbol{\epsilon}_T\|^2)^{1/2}\\ &\leq (1-\eta\lambda_{\mathrm{WD}})M + M\\ &= (2-\eta\lambda_{\mathrm{WD}})M.\end{aligned}$$

We have now bounded both components of the homogeneous state, for every unit initial displacement and zero initial momentum.

Step 2. Measure growth with a quadratic form.

Write $\boldsymbol{z}_t := (\boldsymbol{\epsilon}_t, \boldsymbol{p}_t)^\top$. The two updates in Step 1 give

$$\begin{aligned}\boldsymbol{z}_{t+1} &= \widehat{\boldsymbol{F}}_t\boldsymbol{z}_t,\\ \widehat{\boldsymbol{F}}_t &:= \begin{pmatrix}(1-\eta\lambda_{\mathrm{WD}})\boldsymbol{I} - \alpha(\boldsymbol{P}^{-1}\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I}) & -\beta_1\boldsymbol{I}\\ \alpha(\boldsymbol{P}^{-1}\boldsymbol{H}_{\mathcal{B}_t} + \lambda_{\mathsf{P}}\boldsymbol{I}) & \beta_1\boldsymbol{I}\end{pmatrix}.\end{aligned} \tag{86}$$

This state is related to the existing position–momentum state of (67) by $\boldsymbol{x}_t = \boldsymbol{P}^{1/2}\boldsymbol{\epsilon}_t$ and $\boldsymbol{n}_t = \eta^{-1}\boldsymbol{P}^{1/2}\boldsymbol{p}_t$. This change is invertible for every $\eta > 0$. The moment map for $\widehat{\boldsymbol{F}}_t$ thus has spectral radius ρ, by the coordinate argument in Appendix G.6, which also covers $\eta\lambda_{\mathrm{WD}} = 1$.

We seek a matrix $\boldsymbol{W} \succeq 0$, $\boldsymbol{W} \neq 0$, such that

$$\mathbb{E}[\widehat{\boldsymbol{F}}_t^\top \boldsymbol{W} \widehat{\boldsymbol{F}}_t] = \rho \boldsymbol{W}.$$

This identity has a direct purpose: it makes the expected value of $\boldsymbol{z}_t^\top \boldsymbol{W} \boldsymbol{z}_t$ multiply by ρ after one step. Such a matrix exists by the Perron–Frobenius theorem for PSD-preserving maps (Lemma G.16), applied to the adjoint moment map $\boldsymbol{W} \mapsto \mathbb{E}[\widehat{\boldsymbol{F}}_t^\top \boldsymbol{W} \widehat{\boldsymbol{F}}_t]$. The adjoint has the same spectral radius as the moment map and preserves PSD matrices.

Let $\boldsymbol{W}_{11}$ and $\boldsymbol{W}_{22}$ be the two diagonal blocks of $\boldsymbol{W}$, corresponding to displacement and scaled momentum. We next bound their relative size, so that Step 1 can control this quadratic form. Since $\boldsymbol{W} \succeq 0$, the Euclidean triangle inequality gives, for any vectors $\boldsymbol{u}, \boldsymbol{v}$,

$$\begin{aligned}\left[\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{pmatrix}^\top \boldsymbol{W} \begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{pmatrix}\right]^{1/2} &= \left\| \boldsymbol{W}^{1/2} \begin{pmatrix}\boldsymbol{u}\\ 0\end{pmatrix} + \boldsymbol{W}^{1/2} \begin{pmatrix}0\\ \boldsymbol{v}\end{pmatrix} \right\| \\ &\leq (\boldsymbol{u}^\top \boldsymbol{W}_{11} \boldsymbol{u})^{1/2} + (\boldsymbol{v}^\top \boldsymbol{W}_{22} \boldsymbol{v})^{1/2}.\end{aligned}$$

The momentum column of $\widehat{\boldsymbol{F}}_t$ is deterministic: $\widehat{\boldsymbol{F}}_t(0, \boldsymbol{v})^\top = \beta_1(-\boldsymbol{v}, \boldsymbol{v})^\top$. Testing the eigenmatrix identity on $(0, \boldsymbol{v})$ therefore gives

$$\begin{aligned}\rho\, \boldsymbol{v}^\top \boldsymbol{W}_{22} \boldsymbol{v} &= \begin{pmatrix}0\\ \boldsymbol{v}\end{pmatrix}^\top \mathbb{E}[\widehat{\boldsymbol{F}}_t^\top \boldsymbol{W} \widehat{\boldsymbol{F}}_t] \begin{pmatrix}0\\ \boldsymbol{v}\end{pmatrix} \\ &= \beta_1^2 \begin{pmatrix}-\boldsymbol{v}\\ \boldsymbol{v}\end{pmatrix}^\top \boldsymbol{W} \begin{pmatrix}-\boldsymbol{v}\\ \boldsymbol{v}\end{pmatrix} \\ &\leq \beta_1^2 \left[(\boldsymbol{v}^\top \boldsymbol{W}_{11} \boldsymbol{v})^{1/2} + (\boldsymbol{v}^\top \boldsymbol{W}_{22} \boldsymbol{v})^{1/2}\right]^2.\end{aligned}$$

Taking square roots and moving the $\boldsymbol{W}_{22}$ term to the left gives, for every $\boldsymbol{v}$,

$$(\sqrt{\rho} - \beta_1)(\boldsymbol{v}^\top \boldsymbol{W}_{22} \boldsymbol{v})^{1/2} \leq \beta_1 (\boldsymbol{v}^\top \boldsymbol{W}_{11} \boldsymbol{v})^{1/2}.$$

The displacement block $\boldsymbol{W}_{11}$ is nonzero. Otherwise $\sqrt{\rho} - \beta_1 > 0$ would force $\boldsymbol{W}_{22} = 0$. The triangle inequality above would then give $\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z} = 0$ for every $\boldsymbol{z}$, contradicting $\boldsymbol{W} \neq 0$.

Rescale $\boldsymbol{W}$ so that $\|\boldsymbol{W}_{11}\| = 1$, where the norm is the spectral norm. The eigenmatrix identity is unchanged by this positive scalar rescaling. Dividing the preceding inequality by $\sqrt{\rho} - \beta_1 > 0$ and maximizing over unit $\boldsymbol{v}$ gives

$$\begin{aligned}\sqrt{\|\boldsymbol{W}_{22}\|} &= \max_{\|\boldsymbol{v}\|=1} (\boldsymbol{v}^\top \boldsymbol{W}_{22} \boldsymbol{v})^{1/2} \\ &\leq \frac{\beta_1}{\sqrt{\rho} - \beta_1} \max_{\|\boldsymbol{v}\|=1} (\boldsymbol{v}^\top \boldsymbol{W}_{11} \boldsymbol{v})^{1/2} \\ &= \frac{\beta_1}{\sqrt{\rho} - \beta_1} \leq \frac{\beta_1}{1 - \beta_1}.\end{aligned}$$

The last line uses $\|\boldsymbol{W}_{11}\| = 1$ and $\rho \geq 1$. Applying the same triangle inequality to an arbitrary state $\boldsymbol{z} = (\boldsymbol{\epsilon}, \boldsymbol{p})^\top$ now yields

$$\begin{aligned}(\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z})^{1/2} &\leq (\boldsymbol{\epsilon}^\top \boldsymbol{W}_{11} \boldsymbol{\epsilon})^{1/2} + (\boldsymbol{p}^\top \boldsymbol{W}_{22} \boldsymbol{p})^{1/2} \\ &\leq \|\boldsymbol{\epsilon}\| + \sqrt{\|\boldsymbol{W}_{22}\|}\, \|\boldsymbol{p}\| \\ &\leq \|\boldsymbol{\epsilon}\| + \frac{\beta_1}{1 - \beta_1} \|\boldsymbol{p}\|.\end{aligned} \tag{87}$$

Step 3. Compare exact growth with the detector bounds.

Choose a unit top eigenvector $\boldsymbol{v}$ of $\boldsymbol{W}_{11}$. Our normalization gives $\boldsymbol{v}^\top \boldsymbol{W}_{11}\boldsymbol{v} = 1$. Use the homogeneous trajectory from Step 1 with this initial displacement and zero initial momentum. Its initial quadratic form is

$$\boldsymbol{z}_0^\top \boldsymbol{W} \boldsymbol{z}_0 = \begin{pmatrix} \boldsymbol{v} \\ 0 \end{pmatrix}^\top \boldsymbol{W} \begin{pmatrix} \boldsymbol{v} \\ 0 \end{pmatrix} = \boldsymbol{v}^\top \boldsymbol{W}_{11} \boldsymbol{v} = 1.$$

This choice is allowed because acceptance covers every unit initial displacement.

The batch at time t is independent of $\boldsymbol{z}_t$. The tower property and the eigenmatrix identity give

$$\begin{aligned} \mathbb{E}[\boldsymbol{z}_{t+1}^\top \boldsymbol{W} \boldsymbol{z}_{t+1}] &= \mathbb{E}[\boldsymbol{z}_t^\top \widehat{\boldsymbol{F}}_t^\top \boldsymbol{W} \widehat{\boldsymbol{F}}_t \boldsymbol{z}_t] \\ &= \mathbb{E}\left[\boldsymbol{z}_t^\top \mathbb{E}[\widehat{\boldsymbol{F}}_t^\top \boldsymbol{W} \widehat{\boldsymbol{F}}_t \mid \boldsymbol{z}_t] \boldsymbol{z}_t\right] \\ &= \rho\, \mathbb{E}[\boldsymbol{z}_t^\top \boldsymbol{W} \boldsymbol{z}_t]. \end{aligned}$$

Iterating from the initial value one gives $\mathbb{E}[\boldsymbol{z}_T^\top \boldsymbol{W} \boldsymbol{z}_T] = \rho^T$. On the other hand, (87), Minkowski's inequality and the bounds from Step 1 imply

$$\begin{aligned} \rho^{T/2} &= \left(\mathbb{E}[\boldsymbol{z}_T^\top \boldsymbol{W} \boldsymbol{z}_T]\right)^{1/2} \\ &\le \left[\mathbb{E}\left(\|\boldsymbol{\epsilon}_T\| + \frac{\beta_1}{1-\beta_1}\|\boldsymbol{p}_T\|\right)^2\right]^{1/2} \\ &\le (\mathbb{E}\,\|\boldsymbol{\epsilon}_T\|^2)^{1/2} + \frac{\beta_1}{1-\beta_1}(\mathbb{E}\,\|\boldsymbol{p}_T\|^2)^{1/2} \\ &\le M + \frac{\beta_1}{1-\beta_1}(2 - \eta\lambda_{\mathrm{WD}})M \\ &= \frac{M\left[1 - \beta_1 + 2\beta_1 - \beta_1\eta\lambda_{\mathrm{WD}}\right]}{1-\beta_1} \\ &= \frac{M(1+b)}{1-\beta_1}. \end{aligned}$$

Finally, $b = \beta_1(1 - \eta\lambda_{\mathrm{WD}}) = \beta_1|1 - \eta\lambda_{\mathrm{WD}}|$ and $b \le \beta_1$ on the search window. Taking the power $2/T$,

$$\begin{aligned} \rho &\le \left[\frac{M(1+b)}{1-\beta_1}\right]^{2/T} \\ &= \left[\frac{M(1+\beta_1|1-\eta\lambda_{\mathrm{WD}}|)}{1-\beta_1}\right]^{2/T} \\ &\le \left[\frac{M(1+\beta_1)}{1-\beta_1}\right]^{2/T} = \rho_T. \end{aligned}$$

This completes the proof of Lemma G.9. □

Remark G.10 (Fixed weighted displacement norms). For any fixed $\boldsymbol{W}_0 \succ 0$, Lemma G.9 holds with the same constant when both the unit initialization and the displacement cutoff use $\|\boldsymbol{\epsilon}\|_{\boldsymbol{W}_0}^2 = \boldsymbol{\epsilon}^\top \boldsymbol{W}_0 \boldsymbol{\epsilon}$: replacing $(\boldsymbol{\epsilon}, \boldsymbol{p})$ by $(\boldsymbol{W}_0^{1/2}\boldsymbol{\epsilon}, \boldsymbol{W}_0^{1/2}\boldsymbol{p})$ preserves the scalar β_1 and $1 - \eta\lambda_{\mathrm{WD}}$ blocks in the proof and changes the moment map by a similarity. This includes the fixed preconditioner norm used by the implemented detector.

G.4.2 Bounding the Inverse LR

Proof of Proposition G.6. We apply the quantitative stability bound from Proposition G.2 to the compensated system. Its critical test matrix then converts the quotient difference into the desired inverse-rate difference.

Fix an accepted LR $\eta \in \mathcal{I}$. If $\eta \le \eta_\star^{\mathsf{P}}$, then $1/\eta \ge 1/\eta_\star^{\mathsf{P}}$, as claimed. Assume $\eta > \eta_\star^{\mathsf{P}}$ from now on. Recall the candidate and critical coefficients,

$$\begin{aligned} b &= \beta_1(1 - \eta\lambda_{\mathrm{WD}}), & \alpha &= (1-\beta_1)\eta, \\ b_\star^{\mathsf{P}} &= \beta_1(1 - \eta_\star^{\mathsf{P}}\lambda_{\mathrm{WD}}), & \alpha_\star^{\mathsf{P}} &= (1-\beta_1)\eta_\star^{\mathsf{P}}. \end{aligned}$$

The search window and $\eta > \eta_\star^{\mathsf{P}}$ imply

$$0 \le b \le b_\star^{\mathsf{P}} \le \beta_1 < 1, \qquad \alpha > \alpha_\star^{\mathsf{P}} > 0.$$

The matrix $\boldsymbol{X}_\star^{\mathsf{P}} \succeq 0$ maximizes the compensated quotient at its critical rate, with $\left\|\boldsymbol{X}_\star^{\mathsf{P}}\right\|_F = 1$. Its curvature is $\mu_\star^{\mathsf{P}} = \operatorname{tr}(\boldsymbol{A}_{\mathsf{P}}(\boldsymbol{X}_\star^{\mathsf{P}})^2) > 0$.

Step 1. Use acceptance to bound the quotient.

Lemma G.9 gives $\rho(\mathcal{T}_\eta^{\mathsf{P}}) \le \rho_T$. Apply Proposition G.2 with the compensated mean $\boldsymbol{C} = \boldsymbol{A}_{\mathsf{P}}$, its full noise $\boldsymbol{\Delta}_t$, radius bound $R = \rho_T$, and test matrix $\boldsymbol{X} = \boldsymbol{X}_\star^{\mathsf{P}}$. This gives

$$q_{\boldsymbol{A}_{\mathsf{P}},b}(\boldsymbol{X}_\star^{\mathsf{P}}) - \frac{1-b}{\alpha} \le (\rho_T - 1)\left[\frac{2b}{(\rho_T + b)\alpha} + \frac{(1 - b/\rho_T)^2}{2\mu_\star^{\mathsf{P}}\alpha^2}\right]. \tag{88}$$

Step 2. Convert the quotient into an inverse-rate difference.

At the critical rate, the quotient maximum equals the normalized inverse LR. Recall (72) and the AGM parameter curve:

$$q_{\boldsymbol{A}_{\mathsf{P}},b_\star^{\mathsf{P}}}(\boldsymbol{X}_\star^{\mathsf{P}}) = \frac{1-b_\star^{\mathsf{P}}}{\alpha_\star^{\mathsf{P}}} = \frac{1}{\eta_\star^{\mathsf{P}}} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1},$$
$$\frac{1-b}{\alpha} = \frac{1-\beta_1+\beta_1\eta\lambda_{\mathrm{WD}}}{(1-\beta_1)\eta} = \frac{1}{\eta} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}.$$

Subtracting cancels the same weight-decay offset:

$$\begin{aligned} q_{\boldsymbol{A}_{\mathsf{P}},b_\star^{\mathsf{P}}}(\boldsymbol{X}_\star^{\mathsf{P}}) - \frac{1-b}{\alpha} &= \left(\frac{1}{\eta_\star^{\mathsf{P}}} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}\right) - \left(\frac{1}{\eta} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1-\beta_1}\right) \\ &= \frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta}. \end{aligned} \tag{89}$$

Because $b \le b_\star^{\mathsf{P}}$, the pointwise momentum inequality (65) gives $q_{\boldsymbol{A}_{\mathsf{P}},b_\star^{\mathsf{P}}}(\boldsymbol{X}_\star^{\mathsf{P}}) \le q_{\boldsymbol{A}_{\mathsf{P}},b}(\boldsymbol{X}_\star^{\mathsf{P}})$. Hence

$$\begin{aligned} \frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta} &= q_{\boldsymbol{A}_{\mathsf{P}},b_\star^{\mathsf{P}}}(\boldsymbol{X}_\star^{\mathsf{P}}) - \frac{1-b}{\alpha} \\ &\le q_{\boldsymbol{A}_{\mathsf{P}},b}(\boldsymbol{X}_\star^{\mathsf{P}}) - \frac{1-b}{\alpha} \\ &\le (\rho_T - 1)\left[\frac{2b}{(\rho_T+b)\alpha} + \frac{1}{2\mu_\star^{\mathsf{P}}}\left(\frac{1-b/\rho_T}{\alpha}\right)^2\right]. \end{aligned}$$

The last line uses (88). It remains to replace the candidate coefficients by the fixed critical coefficients.

Step 3. Obtain the uniform bound H_T.

For the first term, $0 \le b \le b_\star^{\mathsf{P}}$ gives

$$\frac{b_\star^{\mathsf{P}}}{\rho_T + b_\star^{\mathsf{P}}} - \frac{b}{\rho_T + b} = \frac{\rho_T(b_\star^{\mathsf{P}} - b)}{(\rho_T + b_\star^{\mathsf{P}})(\rho_T + b)} \ge 0.$$

Combining this with $\alpha > \alpha_\star^{\mathsf{P}} > 0$,

$$\frac{b}{(\rho_T+b)\alpha} \le \frac{b_\star^{\mathsf{P}}}{(\rho_T + b_\star^{\mathsf{P}})\alpha} \le \frac{b_\star^{\mathsf{P}}}{(\rho_T + b_\star^{\mathsf{P}})\alpha_\star^{\mathsf{P}}}.$$

For the second term, we compare the entire ratio $(1 - b/\rho_T)/\alpha$. Substituting the AGM expressions for b and α,

$$\begin{aligned}\frac{1 - b/\rho_T}{\alpha} &= \frac{1 - \beta_1(1 - \eta\lambda_{\mathrm{WD}})/\rho_T}{(1-\beta_1)\eta} \\ &= \frac{1}{1-\beta_1}\left[\frac{1-\beta_1/\rho_T}{\eta} + \frac{\beta_1\lambda_{\mathrm{WD}}}{\rho_T}\right] \\ &\le \frac{1}{1-\beta_1}\left[\frac{1-\beta_1/\rho_T}{\eta_\star^{\mathsf{P}}} + \frac{\beta_1\lambda_{\mathrm{WD}}}{\rho_T}\right] \\ &= \frac{1 - \beta_1(1 - \eta_\star^{\mathsf{P}}\lambda_{\mathrm{WD}})/\rho_T}{(1-\beta_1)\eta_\star^{\mathsf{P}}} = \frac{1 - b_\star^{\mathsf{P}}/\rho_T}{\alpha_\star^{\mathsf{P}}}.\end{aligned}$$

The inequality uses $\eta > \eta_\star^{\mathsf{P}}$ and $1 - \beta_1/\rho_T > 0$. Both ratios are positive, so the inequality remains valid after squaring. Substituting these two comparisons gives exactly the definition of H_T from (56):

$$\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta} \le (\rho_T - 1)\left[\frac{2b_\star^{\mathsf{P}}}{(\rho_T + b_\star^{\mathsf{P}})\alpha_\star^{\mathsf{P}}} + \frac{(1 - b_\star^{\mathsf{P}}/\rho_T)^2}{2\mu_\star^{\mathsf{P}}(\alpha_\star^{\mathsf{P}})^2}\right] = H_T.$$

For $\beta_1 = 0$, $b = 0$ and $\alpha = \eta$, so the same quotient bound (88) also gives the direct LR bound:

$$\begin{aligned}\frac{1}{\eta_\star^{\mathsf{P}}} - \frac{1}{\eta} &\le \frac{\rho_T - 1}{2\mu_\star^{\mathsf{P}}\eta^2}, \\ \eta^2 - \eta_\star^{\mathsf{P}}\eta &\le \frac{(\rho_T - 1)\eta_\star^{\mathsf{P}}}{2\mu_\star^{\mathsf{P}}}, \\ \eta &\le \frac{\eta_\star^{\mathsf{P}}}{2}\left[1 + \sqrt{1 + \frac{2(\rho_T - 1)}{\mu_\star^{\mathsf{P}}\eta_\star^{\mathsf{P}}}}\right] \\ &\le \eta_\star^{\mathsf{P}} + \frac{\rho_T - 1}{2\mu_\star^{\mathsf{P}}}.\end{aligned}$$

The second line multiplies the first by $\eta_\star^{\mathsf{P}}\eta^2 > 0$. The third takes the positive root of the resulting quadratic, and the last uses $\sqrt{1+x} \le 1 + x/2$ for $x \ge 0$. This completes the proof of Proposition G.6. □

For a scalar system without momentum ($\beta_1 = 0$), with compensated mean a and $\eta > \eta_\star^{\mathsf{P}}$, the quantitative bound (64) is an equality when R is the actual radius $\rho(\mathcal{T}_\eta^{\mathsf{P}}) = \mathbb{E}(1 - \eta(a + \boldsymbol{\Delta}_t))^2$. The inverse LR bound then loses only the gap between this radius and ρ_T in Lemma G.9, and the final relaxation $1/\eta \le 1/\eta_\star^{\mathsf{P}}$. With momentum, the momentum comparisons can introduce a first-order loss when the noise is small.

G.4.3 Choosing the Cutoff

At $\eta_\star^{\mathsf{P}}$, $\rho(\mathcal{T}_{\eta_\star^{\mathsf{P}}}^{\mathsf{P}}) = 1$ by Corollary G.4, but this does not ensure bounded affine moments or acceptance at a finite cutoff. For example, without noise, forcing or weight decay, the top eigendirection at $\alpha\lambda_{\max}(\boldsymbol{A}_{\mathsf{P}}) = 2(1 + \beta_1)$ obeys $x_{t+1} = -(1 + \beta_1)x_t - \beta_1 x_{t-1}$. From a unit displacement with zero momentum ($x_{-1} = x_0 = 1$), its solution is

$$\begin{aligned}x_t &= \frac{1 + \beta_1}{1 - \beta_1}(-1)^t - \frac{2\beta_1}{1 - \beta_1}(-\beta_1)^t, \\ |x_t| &\longrightarrow \frac{1 + \beta_1}{1 - \beta_1},\end{aligned}$$

At $\beta_1 = 0.9$, the limiting amplitude is 19, so any smaller cutoff M rejects $\eta_\star^{\mathsf{P}}$ for sufficiently large T despite bounded displacement. The affine detector may also fail its cutoff because of batch-gradient forcing, weight-decay drift or checkpoint momentum.

Acceptance at $\eta_\star^{\mathsf{P}}$ gives $\widehat{\eta}_\star(M, T) \ge \eta_\star^{\mathsf{P}}$, which supplies the upper bound in Theorem D.16. The lower bound only needs a nonempty accepted set and $\eta_\star^{\mathsf{P}} \in \mathcal{I}$, which ensures nonnegative effective momentum at $\eta_\star^{\mathsf{P}}$. Proposition G.11 gives a sufficient cutoff that grows polynomially with T.

We now fix $\eta = \eta_\star^{\mathsf{P}}$ and return to the original affine simulation, with its prescribed checkpoint momentum and forcing. The state below is the state of this single run, not a difference of coupled runs or a rescaled trajectory. Use the original-coordinate state from the proof of Lemma G.9, $\boldsymbol{z}_t = (\boldsymbol{\epsilon}_t, \eta_\star^{\mathsf{P}} \boldsymbol{P}^{-1} \boldsymbol{m}_t)^\top$. With $\alpha = \eta_\star^{\mathsf{P}}(1-\beta_1)$, it satisfies

$$\begin{aligned} \boldsymbol{z}_{t+1} &= \widehat{\boldsymbol{F}}_t \boldsymbol{z}_t + \boldsymbol{f}_t, \\ \boldsymbol{f}_t &\coloneqq \begin{pmatrix} -\alpha \boldsymbol{P}^{-1} \boldsymbol{g}_{\mathcal{B}_t} - \eta_\star^{\mathsf{P}} \lambda_{\mathrm{WD}} \boldsymbol{w}_{t_0} \\ \alpha \boldsymbol{P}^{-1} \boldsymbol{g}_{\mathcal{B}_t} \end{pmatrix}, \end{aligned} \tag{90}$$

with the homogeneous transition matrix $\widehat{\boldsymbol{F}}_t$ from (86), evaluated at $\eta = \eta_\star^{\mathsf{P}}$. Thus all quantities in this cutoff calculation use the critical LR. Write $\widehat{\boldsymbol{F}}_{t:s} \coloneqq \widehat{\boldsymbol{F}}_{t-1} \cdots \widehat{\boldsymbol{F}}_s$ for the propagator from time s to time t, with $\widehat{\boldsymbol{F}}_{t:t} = \boldsymbol{I}$, and define

$$\begin{aligned} K_T &\coloneqq \max_{0 \le t \le T} \left\| \mathbb{E}[\widehat{\boldsymbol{F}}_{t:0}^\top \widehat{\boldsymbol{F}}_{t:0}] \right\|, \\ R_0 &\coloneqq \left(1 + \left\| \eta_\star^{\mathsf{P}} \boldsymbol{P}^{-1} \boldsymbol{m}_{t_0} \right\|^2\right)^{1/2}, \\ \sigma_f &\coloneqq \left(\mathbb{E} \left\| \boldsymbol{f}_t \right\|^2\right)^{1/2}. \end{aligned} \tag{91}$$

$K_T \ge 1$ is the worst-case second-moment growth of the homogeneous run over the horizon, $\mathbb{E} \left\| \widehat{\boldsymbol{F}}_{t:0} \boldsymbol{z} \right\|^2 \le K_T \left\| \boldsymbol{z} \right\|^2$ for $t \le T$. The quantity R_0 is the norm of the initial state for a unit displacement with the checkpoint momentum. The size of the forcing is

$$\sigma_f^2 = \mathbb{E} \left\| \alpha \boldsymbol{P}^{-1} \boldsymbol{g}_{\mathcal{B}_t} + \eta_\star^{\mathsf{P}} \lambda_{\mathrm{WD}} \boldsymbol{w}_{t_0} \right\|^2 + \alpha^2 \mathbb{E} \left\| \boldsymbol{P}^{-1} \boldsymbol{g}_{\mathcal{B}_t} \right\|^2.$$

Proposition G.11 (An explicit cutoff for acceptance at the compensated critical LR)**.** *Under Assumption D.14, suppose $\eta_\star^{\mathsf{P}} \in \mathcal{I}$. If*

$$M \ge \sqrt{K_T} \left(R_0 + T\sigma_f\right), \tag{92}$$

then $\eta_\star^{\mathsf{P}}$ is accepted by the detector (48), that is, Assumption D.15 holds.

Proof of Proposition G.11. We use K_T, the maximal homogeneous second-moment growth over the horizon, to control the initial state and each forcing term separately. Recall that $\left\| \boldsymbol{z}_0 \right\| = R_0$ for every unit initial displacement with the prescribed checkpoint momentum, and $\mathbb{E} \left\| \boldsymbol{f}_j \right\|^2 = \sigma_f^2$. Unrolling the affine recursion (90),

$$\boldsymbol{z}_t = \widehat{\boldsymbol{F}}_{t:0} \boldsymbol{z}_0 + \sum_{j=0}^{t-1} \widehat{\boldsymbol{F}}_{t:j+1} \boldsymbol{f}_j.$$

The propagator $\widehat{\boldsymbol{F}}_{t:j+1}$ depends on the batches $j+1, \dots, t-1$ and $\boldsymbol{f}_j$ on batch j, so the two are independent. By the tower property,

$$\begin{aligned} \mathbb{E} \left\| \widehat{\boldsymbol{F}}_{t:j+1} \boldsymbol{f}_j \right\|^2 &= \mathbb{E}\left[\boldsymbol{f}_j^\top \, \mathbb{E}[\widehat{\boldsymbol{F}}_{t:j+1}^\top \widehat{\boldsymbol{F}}_{t:j+1}] \, \boldsymbol{f}_j \right] \\ &\le \left\| \mathbb{E}[\widehat{\boldsymbol{F}}_{t:j+1}^\top \widehat{\boldsymbol{F}}_{t:j+1}] \right\| \sigma_f^2 \\ &= \left\| \mathbb{E}[\widehat{\boldsymbol{F}}_{t-j-1:0}^\top \widehat{\boldsymbol{F}}_{t-j-1:0}] \right\| \sigma_f^2 \\ &\le K_T \sigma_f^2, \end{aligned}$$

where the equality holds because the batches are i.i.d., so that $\widehat{\boldsymbol{F}}_{t:j+1}$ has the law of $\widehat{\boldsymbol{F}}_{t-j-1:0}$. Likewise $\mathbb{E} \left\| \widehat{\boldsymbol{F}}_{t:0} \boldsymbol{z}_0 \right\|^2 \le K_T \left\| \boldsymbol{z}_0 \right\|^2 = K_T R_0^2$ for a unit displacement. Minkowski's inequality in L^2 gives, for $t \le T$,

$$\begin{aligned} \left(\mathbb{E} \left\| \boldsymbol{z}_t \right\|^2\right)^{1/2} &\le \left(\mathbb{E} \left\| \widehat{\boldsymbol{F}}_{t:0} \boldsymbol{z}_0 \right\|^2\right)^{1/2} + \sum_{j=0}^{t-1} \left(\mathbb{E} \left\| \widehat{\boldsymbol{F}}_{t:j+1} \boldsymbol{f}_j \right\|^2\right)^{1/2} \\ &\le \sqrt{K_T} \left(R_0 + t\sigma_f\right) \\ &\le M. \end{aligned}$$

Since $\left\| \boldsymbol{\epsilon}_t \right\| \le \left\| \boldsymbol{z}_t \right\|$, $\mathbb{E} \left\| \boldsymbol{\epsilon}_t \right\|^2 \le M^2$ for all $0 \le t \le T$ and every unit $\boldsymbol{\epsilon}_0$, which is the acceptance of $\eta_\star^{\mathsf{P}}$. This completes the proof of Proposition G.11. □

The condition is not a smallness condition on the noise: K_T, R_0 and σ_f enter with their actual values, and K_T is finite for every T. To control its growth, write

$$\mathcal{U}(\boldsymbol{S}) \coloneqq \mathbb{E}[\widehat{\boldsymbol{F}}_t \boldsymbol{S} \widehat{\boldsymbol{F}}_t^\top], \qquad \boldsymbol{S} \in \mathbb{S}_{2d},$$

for the homogeneous moment map at $\eta_\star^{\mathsf{P}}$. The critical-rate result (Corollary G.4) and the change of coordinates above give $\rho(\mathcal{U}) = 1$. The Jordan decomposition therefore bounds the powers of its adjoint, applied to any fixed matrix, by a polynomial. In particular, for some fixed $C' > 1$ and integer $0 \le k < \dim \mathbb{S}_{2d}$,

$$\left\| (\mathcal{U}^*)^t(\boldsymbol{I}_{2d}) \right\|_F \le C'(1+t)^k.$$

Iterating the adjoint map uses the same independence across steps as the forward second-moment recursion. Consequently,

$$\begin{aligned}
\mathbb{E}[\widehat{\boldsymbol{F}}_{t:0}^\top \widehat{\boldsymbol{F}}_{t:0}] &= (\mathcal{U}^*)^t(\boldsymbol{I}_{2d}), && \text{iterate the adjoint moment map} \\
\left\| \mathbb{E}[\widehat{\boldsymbol{F}}_{t:0}^\top \widehat{\boldsymbol{F}}_{t:0}] \right\| &\le \left\| (\mathcal{U}^*)^t(\boldsymbol{I}_{2d}) \right\|_F \\
&\le C'(1+t)^k, && \text{Jordan power bound} \\
K_T &\le C'(1+T)^k. && \text{maximize over } 0 \le t \le T
\end{aligned}$$

The resulting cutoff satisfies

$$\begin{aligned}
M_T &\coloneqq \sqrt{C'(1+T)^k}\,(R_0 + T\sigma_f) \\
&\ge \sqrt{K_T}\,(R_0 + T\sigma_f), && \text{the sufficient cutoff above} \\
\frac{1+\log M_T}{T} &= O\!\left(\frac{\log T}{T}\right) \longrightarrow 0. && \text{fixed model and polynomial cutoff}
\end{aligned}$$

Thus $M_T > 1$ is admissible at every horizon and yields $H_T = O(\log T/T)$ for a fixed model. The factor $T\sigma_f$ comes from Minkowski's inequality and does not assert linear growth of the actual run.

G.5 Proofs of the Stability Criterion

We prove the exact criterion, its quantitative form, and the AGM critical-rate statement from Appendix G.1. For the exact and quantitative criteria, we use the generic recurrence

$$\boldsymbol{x}_{t+1} = \big[(1+b)\boldsymbol{I} - \alpha(\boldsymbol{C} + \boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t - b\boldsymbol{x}_{t-1}, \qquad \boldsymbol{C} \succ 0, \quad -1 < b < 1, \quad \alpha > 0.$$

The AGM proof also treats the remaining LRs, for which $b \le -1$. Recall that $\mathcal{T}(\boldsymbol{S}) = \mathbb{E}[\boldsymbol{F}_t \boldsymbol{S} \boldsymbol{F}_t^\top]$ is its second-moment map, with the history transition $\boldsymbol{F}_t$ from (62). The quotient $q_{\boldsymbol{C},b}$ and its maximum $r_{\boldsymbol{C}}(b)$ are recalled in Appendix G.1.

G.5.1 Proof of the Exact Criterion

A general random linear recurrence Consider the random linear recurrence

$$\boldsymbol{y}_{t+1} = \boldsymbol{E}_t \boldsymbol{y}_t, \qquad \boldsymbol{y}_t \in \mathbb{R}^m, \tag{93}$$

where the matrices $\boldsymbol{E}_t$ are i.i.d. with finite second moments and the initial state $\boldsymbol{y}_0$ is deterministic. We seek a criterion for $\mathbb{E}\,\|\boldsymbol{y}_t\|^2 \to 0$ for every initial state. Separate the mean matrix from its centered perturbation,

$$\begin{aligned}
\boldsymbol{F} &\coloneqq \mathbb{E}\boldsymbol{E}_t, \qquad \boldsymbol{N}_t \coloneqq \boldsymbol{E}_t - \boldsymbol{F}, \\
\mathcal{T}_0(\boldsymbol{S}) &\coloneqq \boldsymbol{F}\boldsymbol{S}\boldsymbol{F}^\top, \qquad \mathcal{N}(\boldsymbol{S}) \coloneqq \mathbb{E}[\boldsymbol{N}_t \boldsymbol{S} \boldsymbol{N}_t^\top].
\end{aligned} \tag{94}$$

Both $\mathcal{T}_0$ and $\mathcal{N}$ preserve positive semidefinite matrices. The second moment $\boldsymbol{\Sigma}_t \coloneqq \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}_t^\top]$ evolves as

$$\begin{aligned}
\boldsymbol{\Sigma}_{t+1} &= \mathbb{E}[(\boldsymbol{F} + \boldsymbol{N}_t)\boldsymbol{\Sigma}_t(\boldsymbol{F} + \boldsymbol{N}_t)^\top] \\
&= \mathcal{T}_0(\boldsymbol{\Sigma}_t) + \mathcal{N}(\boldsymbol{\Sigma}_t). && \mathbb{E}\boldsymbol{N}_t = 0
\end{aligned} \tag{95}$$

The first equality uses the independence of $\boldsymbol{y}_t$ and $\boldsymbol{E}_t$. Write $\mathcal{T} \coloneqq \mathcal{T}_0 + \mathcal{N}$ for this second-moment map.

Suppose first that the mean dynamics are stable, $\rho(\boldsymbol{F}) < 1$. A perturbation with second moment $\boldsymbol{Q}$ injected once produces the later second moments $\boldsymbol{F}^j \boldsymbol{Q} (\boldsymbol{F}^\top)^j$. Their sum is

$$\mathcal{K}(\boldsymbol{Q}) := \sum_{j \geq 0} \boldsymbol{F}^j \boldsymbol{Q} (\boldsymbol{F}^\top)^j, \qquad \mathcal{K} = (\mathcal{I} - \mathcal{T}_0)^{-1}. \tag{96}$$

Here $\mathcal{I}$ is the identity operator on symmetric matrices. The series converges and preserves PSD matrices. Thus $\mathcal{K}(\boldsymbol{Q})$ is the cumulative response of the mean dynamics to one injection, and $\mathcal{K}\mathcal{N}$ first forms the second moment of a noise term and then accumulates its response. The next proposition characterizes stability through this composition.

Proposition G.12 (Feedback criterion)**.** *The recurrence* (93) *is strictly mean-square stable if and only if* $\rho(\boldsymbol{F}) < 1$ *and* $\rho(\mathcal{K}\mathcal{N}) < 1$.

Proof of Proposition G.12. We transfer a Lyapunov inequality between $\mathcal{T}$ and $\mathcal{K}\mathcal{N}$. The inverse $\mathcal{K}$ accounts for the mean dynamics in this transfer.

Step 1. Separate the mean dynamics.

Strict mean square stability implies $\rho(\boldsymbol{F}) < 1$ by the second-moment characterization (Lemma G.14). Assume $\rho(\boldsymbol{F}) < 1$ from now on, so $\mathcal{K}$ exists. Then

$$\begin{aligned} (\mathcal{I} - \mathcal{T}_0)(\mathcal{I} - \mathcal{K}\mathcal{N}) &= \mathcal{I} - \mathcal{T}_0 - \mathcal{N} && (\mathcal{I} - \mathcal{T}_0)\mathcal{K} = \mathcal{I} \\ &= \mathcal{I} - \mathcal{T}. && \mathcal{T} = \mathcal{T}_0 + \mathcal{N} \end{aligned}$$

Step 2. Necessity.

If $\rho(\mathcal{T}) < 1$, its Neumann series defines

$$\boldsymbol{W} := (\mathcal{I} - \mathcal{T})^{-1}(\boldsymbol{I}_m) = \sum_{j \geq 0} \mathcal{T}^j(\boldsymbol{I}_m) \succeq \boldsymbol{I}_m.$$

The factorization in Step 1 gives

$$\begin{aligned} \boldsymbol{W} - \mathcal{K}\mathcal{N}(\boldsymbol{W}) &= \mathcal{K}\big(\boldsymbol{W} - \mathcal{T}(\boldsymbol{W})\big) \\ &= \mathcal{K}(\boldsymbol{I}_m) \succeq \boldsymbol{I}_m. \end{aligned}$$

The first equality multiplies the factorization by $\mathcal{K}$. The second line uses the Neumann series for $\boldsymbol{W}$ and $\mathcal{K}$. The Lyapunov criterion for PSD-preserving maps (Lemma G.15) therefore gives $\rho(\mathcal{K}\mathcal{N}) < 1$.

Step 3. Sufficiency.

Conversely, if $\rho(\mathcal{K}\mathcal{N}) < 1$, put

$$\boldsymbol{W} := (\mathcal{I} - \mathcal{K}\mathcal{N})^{-1}\mathcal{K}(\boldsymbol{I}_m) = \sum_{j \geq 0} (\mathcal{K}\mathcal{N})^j \mathcal{K}(\boldsymbol{I}_m) \succeq \boldsymbol{I}_m.$$

Now the same factorization gives

$$\begin{aligned} \boldsymbol{W} - \mathcal{T}(\boldsymbol{W}) &= (\mathcal{I} - \mathcal{T}_0)\big(\boldsymbol{W} - \mathcal{K}\mathcal{N}(\boldsymbol{W})\big) && \text{Step 1} \\ &= (\mathcal{I} - \mathcal{T}_0)\mathcal{K}(\boldsymbol{I}_m) = \boldsymbol{I}_m. && \text{definition of } \boldsymbol{W} \end{aligned}$$

The Lyapunov criterion gives $\rho(\mathcal{T}) < 1$, which is strict mean-square stability by Lemma G.14. This completes the proof of Proposition G.12. □

The momentum recurrence We now apply the general criterion to the recurrence from (61). Its mean curvature is $\boldsymbol{C} \succ 0$, its parameters satisfy $-1 < b < 1$ and $\alpha > 0$, and its symmetric noise $\boldsymbol{\Delta}_t$ is i.i.d. with zero mean and finite second moments. The initial pair $(\boldsymbol{x}_0, \boldsymbol{x}_{-1})$ is deterministic. Recall the recurrence and its history state,

$$\boldsymbol{x}_{t+1} = \big[(1+b)\boldsymbol{I} - \alpha(\boldsymbol{C} + \boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t - b\boldsymbol{x}_{t-1}, \qquad \boldsymbol{y}_t := (\boldsymbol{x}_t, \boldsymbol{x}_{t-1})^\top.$$

Write $\boldsymbol{E} \coloneqq (1+b)\boldsymbol{I} - \alpha \boldsymbol{C}$. The random and mean history matrices, and the injection into the position block, are

$$\boldsymbol{F}_t \coloneqq \begin{pmatrix} \boldsymbol{E} - \alpha \boldsymbol{\Delta}_t & -b\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{pmatrix}, \qquad \boldsymbol{F} \coloneqq \mathbb{E}\boldsymbol{F}_t = \begin{pmatrix} \boldsymbol{E} & -b\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{pmatrix}, \qquad \boldsymbol{J} \coloneqq \begin{pmatrix} \boldsymbol{I} \\ 0 \end{pmatrix}. \tag{97}$$

This is the general recurrence with state dimension $m = 2n$ and random matrix $\boldsymbol{E}_t = \boldsymbol{F}_t$. We use $\mathcal{T}_0$, $\mathcal{N}$ and $\mathcal{K}$ for this history system from now on. The matrix $\boldsymbol{J}$ embeds a position vector as $(\boldsymbol{x}, 0)^\top$, while $\boldsymbol{J}^\top \boldsymbol{S} \boldsymbol{J}$ extracts the position block of a history second moment. Let $a_i > 0$ be the eigenvalues of $\boldsymbol{C}$. Recall that $L_{\boldsymbol{C}} = \lambda_{\max}(\boldsymbol{C})$. The characteristic polynomials of $\boldsymbol{F}$ along these eigendirections are $x^2 - (1 + b - \alpha a_i)x + b$. The Jury stability test (Lemma G.17) requires

$$|b| < 1, \qquad \alpha a_i > 0, \qquad 2(1+b) - \alpha a_i > 0.$$

The first two conditions already hold. Thus the mean dynamics are stable if and only if

$$\alpha L_{\boldsymbol{C}} < 2(1+b). \tag{98}$$

The centered perturbation is $\boldsymbol{N}_t = \boldsymbol{F}_t - \boldsymbol{F} = -\alpha \boldsymbol{J} \boldsymbol{\Delta}_t \boldsymbol{J}^\top$, so it acts only on the position. Define its second-moment map on position matrices by

$$\mathcal{V}(\boldsymbol{X}) \coloneqq \mathbb{E}[\boldsymbol{\Delta}_t \boldsymbol{X} \boldsymbol{\Delta}_t], \qquad \boldsymbol{X} \in \mathbb{S}_n. \tag{99}$$

Then the full noise map from (94) is

$$\mathcal{N}(\boldsymbol{S}) = \alpha^2 \boldsymbol{J} \mathcal{V}(\boldsymbol{J}^\top \boldsymbol{S} \boldsymbol{J}) \boldsymbol{J}^\top. \tag{100}$$

Under the mean stability condition $\alpha L_{\boldsymbol{C}} < 2(1+b)$, $\mathcal{K}$ propagates an injection through the mean history matrix $\boldsymbol{F}$ above, as in (96). Since $\mathcal{N}$ uses only the position block, reducing $\mathcal{K}\mathcal{N}$ requires the position response to a position injection. The following lemma computes that response. Recall $c_b \coloneqq (1-b)/(1+b)$ and $k_b \coloneqq b/(1-b^2)$.

Lemma G.13 (Position block of the cumulative response)**.** *Assume (98), and define the operator on* $\mathbb{S}_n$

$$\mathcal{D}(\boldsymbol{X}) \coloneqq \frac{1-b}{\alpha}(\boldsymbol{C}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{C}) - c_b \boldsymbol{C}\boldsymbol{X}\boldsymbol{C} + k_b[\boldsymbol{C}, [\boldsymbol{C}, \boldsymbol{X}]]. \tag{101}$$

Then $\mathcal{D}$ is self-adjoint and positive definite. For every $\boldsymbol{Q} \in \mathbb{S}_n$,

$$\boldsymbol{J}^\top \mathcal{K}(\boldsymbol{J}\boldsymbol{Q}\boldsymbol{J}^\top)\boldsymbol{J} = \alpha^{-2}\mathcal{D}^{-1}(\boldsymbol{Q}). \tag{102}$$

Consequently, $\mathcal{D}^{-1}$ and $\mathcal{D}^{-1}\mathcal{V}$ preserve PSD matrices, and $\mathcal{K}\mathcal{N}$ and $\mathcal{D}^{-1}\mathcal{V}$ have the same nonzero eigenvalues. In particular, their spectral radii are equal.

Proof of Lemma G.13. We solve the Lyapunov equation for a single position injection by eliminating its delayed and cross blocks. This identifies $\mathcal{D}$, while the response series proves that its inverse preserves PSD matrices.

Step 1. Eliminate the cross and delayed blocks.

Fix $\boldsymbol{Q} \in \mathbb{S}_n$. Since (98) gives $\rho(\boldsymbol{F}) < 1$, the cumulative response is

$$\boldsymbol{\Sigma}_{\boldsymbol{Q}} \coloneqq \mathcal{K}(\boldsymbol{J}\boldsymbol{Q}\boldsymbol{J}^\top) = \sum_{j \geq 0} \boldsymbol{F}^j \boldsymbol{J}\boldsymbol{Q}\boldsymbol{J}^\top (\boldsymbol{F}^\top)^j = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Y}^\top & \boldsymbol{W} \end{pmatrix}. \tag{103}$$

Here $\boldsymbol{X}$ is the cumulative position response, $\boldsymbol{W}$ the cumulative delayed-position response, and $\boldsymbol{Y}$ their cross block. These blocks depend linearly on the injection $\boldsymbol{Q}$. Its Lyapunov equation $\boldsymbol{\Sigma}_{\boldsymbol{Q}} = \boldsymbol{J}\boldsymbol{Q}\boldsymbol{J}^\top + \boldsymbol{F}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\boldsymbol{F}^\top$ reads

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{Q} + \boldsymbol{E}\boldsymbol{X}\boldsymbol{E} - b(\boldsymbol{E}\boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{E}) + b^2 \boldsymbol{W}, \\ \boldsymbol{Y} &= \boldsymbol{E}\boldsymbol{X} - b\boldsymbol{Y}^\top, \qquad \boldsymbol{W} = \boldsymbol{X}. \end{aligned}$$

The cross block $\boldsymbol{Y}$ need not be symmetric. Its equation and its transpose give

$$\begin{aligned} \boldsymbol{Y} + b\boldsymbol{Y}^\top &= \boldsymbol{E}\boldsymbol{X}, \qquad \boldsymbol{Y}^\top + b\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{E}, \\ (1-b^2)\boldsymbol{Y} &= \boldsymbol{E}\boldsymbol{X} - b\boldsymbol{X}\boldsymbol{E}, \\ \boldsymbol{E}\boldsymbol{Y} + \boldsymbol{Y}^\top \boldsymbol{E} &= \frac{\boldsymbol{E}^2\boldsymbol{X} + \boldsymbol{X}\boldsymbol{E}^2 - 2b\boldsymbol{E}\boldsymbol{X}\boldsymbol{E}}{1-b^2}. \end{aligned}$$

The second line subtracts b times the transposed equation from the first. The last line substitutes the resulting expressions for $\boldsymbol{Y}$ and $\boldsymbol{Y}^\top$. Substituting these identities and $\boldsymbol{W} = \boldsymbol{X}$ into the first block equation,

$$\begin{aligned}\boldsymbol{Q} &= (1-b^2)\boldsymbol{X} - \boldsymbol{E}\boldsymbol{X}\boldsymbol{E} + b(\boldsymbol{E}\boldsymbol{Y} + \boldsymbol{Y}^\top\boldsymbol{E}) \\ &= (1-b^2)\boldsymbol{X} - \frac{1+b^2}{1-b^2}\boldsymbol{E}\boldsymbol{X}\boldsymbol{E} + \frac{b}{1-b^2}(\boldsymbol{E}^2\boldsymbol{X} + \boldsymbol{X}\boldsymbol{E}^2).\end{aligned}$$

Step 2. Identify the position operator.

Expand $\boldsymbol{E} = (1+b)\boldsymbol{I} - \alpha\boldsymbol{C}$ in the last display. The coefficients of $\boldsymbol{X}$ and $\boldsymbol{C}\boldsymbol{X} + \boldsymbol{X}\boldsymbol{C}$ simplify, respectively, to

$$\begin{aligned}(1-b^2) - \frac{(1+b)^2(1+b^2-2b)}{1-b^2} &= (1-b^2) - (1-b^2) = 0, \\ \frac{\alpha(1+b)(1+b^2-2b)}{1-b^2} &= \alpha(1-b).\end{aligned}$$

The quadratic terms are

$$\begin{aligned}&-\frac{(1+b^2)\alpha^2}{1-b^2}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C} + \frac{b\alpha^2}{1-b^2}(\boldsymbol{C}^2\boldsymbol{X} + \boldsymbol{X}\boldsymbol{C}^2) \\ &= -\frac{(1-b)^2\alpha^2}{1-b^2}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C} + \frac{b\alpha^2}{1-b^2}[\boldsymbol{C},[\boldsymbol{C},\boldsymbol{X}]] \\ &= -\alpha^2 c_b \boldsymbol{C}\boldsymbol{X}\boldsymbol{C} + \alpha^2 k_b [\boldsymbol{C},[\boldsymbol{C},\boldsymbol{X}]].\end{aligned}$$

The second line uses $\boldsymbol{C}^2\boldsymbol{X} + \boldsymbol{X}\boldsymbol{C}^2 = [\boldsymbol{C},[\boldsymbol{C},\boldsymbol{X}]] + 2\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}$. Combining the linear and quadratic terms gives

$$\boldsymbol{Q} = \alpha^2 \mathcal{D}(\boldsymbol{X}). \tag{104}$$

Step 3. Prove positive definiteness.

Each term in (101) is self-adjoint under the Frobenius inner product. For every symmetric $\boldsymbol{X}$, the commutator identity gives

$$\begin{aligned}&c_b \operatorname{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) - k_b \left\|[\boldsymbol{C},\boldsymbol{X}]\right\|_F^2 \\ &\quad = c_b \operatorname{tr}(\boldsymbol{C}^2\boldsymbol{X}^2) - \left(\frac{c_b}{2} + k_b\right)\left\|[\boldsymbol{C},\boldsymbol{X}]\right\|_F^2 \\ &\quad \le c_b \operatorname{tr}(\boldsymbol{C}^2\boldsymbol{X}^2) \le c_b L_{\boldsymbol{C}} \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}^2).\end{aligned} \tag{105}$$

Here $c_b/2 + k_b = (1+b^2)/(2(1-b^2)) > 0$ for every $-1 < b < 1$, and $\boldsymbol{C}^2 \preceq L_{\boldsymbol{C}}\boldsymbol{C}$. Therefore, for $\boldsymbol{X} \neq 0$,

$$\begin{aligned}\langle \boldsymbol{X}, \mathcal{D}\boldsymbol{X}\rangle &\ge \left(\frac{2(1-b)}{\alpha} - c_b L_{\boldsymbol{C}}\right)\operatorname{tr}(\boldsymbol{C}\boldsymbol{X}^2) \\ &> 0.\end{aligned}$$

The first inequality follows from (101) and (105). The second uses $\alpha L_{\boldsymbol{C}} < 2(1+b)$ and $\boldsymbol{C} \succ 0$. Thus $\mathcal{D} \succ 0$. Solving $\boldsymbol{Q} = \alpha^2\mathcal{D}(\boldsymbol{X})$ from (104) gives $\boldsymbol{X} = \alpha^{-2}\mathcal{D}^{-1}(\boldsymbol{Q})$, the position response in (102). If $\boldsymbol{Q} \succeq 0$, every summand of the response series (103) is PSD, so $\boldsymbol{X} \succeq 0$. Hence $\mathcal{D}^{-1}$ preserves PSD matrices. This property follows from the response series, not from positive definiteness of $\mathcal{D}$ as an operator. The composition $\mathcal{D}^{-1}\mathcal{V}$ also preserves PSD matrices.

Step 4. Reduce the spectrum to the position block.

The full noise map injects $\alpha^2\mathcal{V}(\boldsymbol{J}^\top\boldsymbol{S}\boldsymbol{J})$ into the position, and the response of a position injection is $\alpha^{-2}\mathcal{D}^{-1}$. Combining these two identities, (100) and (102), gives

$$\begin{aligned}\mathcal{K}\mathcal{N}(\boldsymbol{S}) &= \alpha^2\mathcal{K}\big(\boldsymbol{J}\mathcal{V}(\boldsymbol{J}^\top\boldsymbol{S}\boldsymbol{J})\boldsymbol{J}^\top\big), \\ \boldsymbol{J}^\top\mathcal{K}\mathcal{N}(\boldsymbol{S})\boldsymbol{J} &= \mathcal{D}^{-1}\mathcal{V}(\boldsymbol{J}^\top\boldsymbol{S}\boldsymbol{J}).\end{aligned} \tag{106}$$

If $\mathcal{KN}(\boldsymbol{S}) = \lambda \boldsymbol{S}$ with $\lambda \neq 0$, compress to $\boldsymbol{X} := \boldsymbol{J}^\top \boldsymbol{S} \boldsymbol{J}$. Then

$$\mathcal{D}^{-1}\mathcal{V}(\boldsymbol{X}) = \lambda \boldsymbol{X}.$$

For $\boldsymbol{S} \neq 0$ this $\boldsymbol{X}$ is nonzero, since otherwise the first line of (106) would give $\lambda \boldsymbol{S} = 0$. Conversely, suppose $\mathcal{D}^{-1}\mathcal{V}(\boldsymbol{X}) = \lambda \boldsymbol{X}$ with $\lambda \neq 0$ and $\boldsymbol{X} \neq 0$. Set $\boldsymbol{S} := \alpha^2 \mathcal{K}(\boldsymbol{J}\mathcal{V}(\boldsymbol{X})\boldsymbol{J}^\top)$. Then

$$\boldsymbol{J}^\top \boldsymbol{S} \boldsymbol{J} = \mathcal{D}^{-1}\mathcal{V}(\boldsymbol{X}) = \lambda \boldsymbol{X} \neq 0,$$
$$\mathcal{KN}(\boldsymbol{S}) = \alpha^2 \mathcal{K}\big(\boldsymbol{J}\mathcal{V}(\lambda \boldsymbol{X})\boldsymbol{J}^\top\big) = \lambda \boldsymbol{S}.$$

The first line uses (102). The second uses (106) and linearity. The nonzero eigenvalues coincide, so the spectral radii are equal. This completes the proof of Lemma G.13. □

We can now express stability through the quotient from (49). For the same mean $\boldsymbol{C}$, noise $\boldsymbol{\Delta}_t$ and fixed momentum b, recall

$$q_{\boldsymbol{C},b}(\boldsymbol{X}) = \frac{\mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) + c_b\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) - k_b\,\|[\boldsymbol{C},\boldsymbol{X}]\|_F^2}{2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)}, \qquad 0 \neq \boldsymbol{X} \succeq 0.$$

Here $c_b = (1-b)/(1+b)$ and $k_b = b/(1-b^2)$, and $r_{\boldsymbol{C}}(b)$ is the maximum over $\boldsymbol{X} \succeq 0$ with $\|\boldsymbol{X}\|_F = 1$. The first numerator term is $\langle \boldsymbol{X}, \mathcal{V}\boldsymbol{X}\rangle$, the noise contribution. The other two come from the mean dynamics through $\mathcal{D}$. Thus

$$\langle \boldsymbol{X}, (\mathcal{D}-\mathcal{V})\boldsymbol{X}\rangle = 2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)\left(\frac{1-b}{\alpha} - q_{\boldsymbol{C},b}(\boldsymbol{X})\right). \tag{107}$$

This identity holds for every nonzero PSD $\boldsymbol{X}$. It follows from $\langle \boldsymbol{X}, \mathcal{V}\boldsymbol{X}\rangle = \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t)$ and $\langle \boldsymbol{X}, [\boldsymbol{C},[\boldsymbol{C},\boldsymbol{X}]]\rangle = \|[\boldsymbol{C},\boldsymbol{X}]\|_F^2$.

Proof of Theorem G.1. The general criterion reduces stability to $\mathcal{D}^{-1}\mathcal{V}$. A PSD eigenmatrix at its spectral radius then reduces the test to the quotient in (107).

Step 1. Reduce to the position operator.

For a unit top eigenvector $\boldsymbol{u}$ of $\boldsymbol{C}$, test the quotient with $\boldsymbol{u}\boldsymbol{u}^\top$:

$$r_{\boldsymbol{C}}(b) \geq q_{\boldsymbol{C},b}(\boldsymbol{u}\boldsymbol{u}^\top) = \frac{c_b L_{\boldsymbol{C}}}{2} + \frac{\mathbb{E}(\boldsymbol{u}^\top \boldsymbol{\Delta}_t \boldsymbol{u})^2}{2L_{\boldsymbol{C}}} \geq \frac{c_b L_{\boldsymbol{C}}}{2} > 0. \tag{108}$$

Thus (63) implies $\alpha L_{\boldsymbol{C}} < 2(1+b)$. Strict mean-square stability also implies this mean stability condition. Under it, Proposition G.12 and Lemma G.13 give

$$\text{strict mean-square stability} \iff \rho(\mathcal{D}^{-1}\mathcal{V}) < 1. \tag{109}$$

Step 2. Test the quadratic form on PSD matrices.

Write $\rho := \rho(\mathcal{D}^{-1}\mathcal{V})$. If $\rho > 0$, the Perron–Frobenius theorem for PSD-preserving maps (Lemma G.16) gives $\boldsymbol{W} \succeq 0$, $\boldsymbol{W} \neq 0$, such that

$$\mathcal{V}(\boldsymbol{W}) = \rho\mathcal{D}(\boldsymbol{W}), \qquad \rho = \frac{\langle \boldsymbol{W}, \mathcal{V}\boldsymbol{W}\rangle}{\langle \boldsymbol{W}, \mathcal{D}\boldsymbol{W}\rangle}.$$

Consequently, positivity of $\langle \boldsymbol{X}, (\mathcal{D}-\mathcal{V})\boldsymbol{X}\rangle$ for every nonzero PSD $\boldsymbol{X}$ implies $\rho < 1$. This implication also holds when $\rho = 0$.

Conversely, suppose $\rho < 1$. We proved in Lemma G.13 that $\mathcal{D}$ is a positive definite self-adjoint operator on $\mathbb{S}_n$, so it has a positive square root $\mathcal{D}^{1/2}$. The operator $\mathcal{V}$ is self-adjoint because the noise matrices are symmetric. The similarity

$$\mathcal{D}^{1/2}(\mathcal{D}^{-1}\mathcal{V})\mathcal{D}^{-1/2} = \mathcal{D}^{-1/2}\mathcal{V}\mathcal{D}^{-1/2}$$

therefore transforms $\mathcal{D}^{-1}\mathcal{V}$ into a self-adjoint operator with the same spectral radius ρ. Its largest eigenvalue is at most ρ. Applying its Rayleigh bound to $\mathcal{D}^{1/2}\boldsymbol{X}$, for every symmetric $\boldsymbol{X} \neq 0$,

$$\begin{aligned}\langle \boldsymbol{X}, \mathcal{V}\boldsymbol{X}\rangle &= \left\langle \mathcal{D}^{1/2}\boldsymbol{X}, (\mathcal{D}^{-1/2}\mathcal{V}\mathcal{D}^{-1/2})\mathcal{D}^{1/2}\boldsymbol{X}\right\rangle \\ &\leq \rho \left\|\mathcal{D}^{1/2}\boldsymbol{X}\right\|_F^2 = \rho\,\langle \boldsymbol{X}, \mathcal{D}\boldsymbol{X}\rangle, && \text{Rayleigh bound} \\ \langle \boldsymbol{X}, (\mathcal{D}-\mathcal{V})\boldsymbol{X}\rangle &\geq (1-\rho)\,\langle \boldsymbol{X}, \mathcal{D}\boldsymbol{X}\rangle > 0. && \rho < 1 \text{ and } \mathcal{D} \succ 0\end{aligned}$$

This proves the equivalence between stability and positivity of the quadratic form on nonzero PSD matrices.

Step 3. Take the maximum of the quotient.

The identity (107) expresses this quadratic form as the positive factor $2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)$ times $(1-b)/\alpha - q_{\boldsymbol{C},b}(\boldsymbol{X})$. Its positivity is therefore equivalent to

$$q_{\boldsymbol{C},b}(\boldsymbol{X}) < \frac{1-b}{\alpha} \qquad \text{for every } \boldsymbol{X} \succeq 0 \text{ with } \|\boldsymbol{X}\|_F = 1.$$

The maximum is attained on this compact set, since its denominator $2\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2)$ is at least $2\lambda_{\min}(\boldsymbol{C}) > 0$. The condition is therefore (63). For fixed b, its right side is independent of α, giving the stated interval. This completes the proof of Theorem G.1. □

G.5.2 Proof of the Quantitative Bound

The proof rescales the recurrence to make it strictly stable and then applies the exact criterion. We first prove the shift identity (Lemma G.3), which compares the rescaled mean with the original one.

Proof of Lemma G.3. The noise term is unchanged because the two systems use the same $\boldsymbol{\Delta}_t$. We compute the other terms separately. First,

$$\begin{aligned}[\boldsymbol{C}+h\boldsymbol{I}, \boldsymbol{X}] &= (\boldsymbol{C}+h\boldsymbol{I})\boldsymbol{X} - \boldsymbol{X}(\boldsymbol{C}+h\boldsymbol{I}) \\ &= \boldsymbol{C}\boldsymbol{X} + h\boldsymbol{X} - \boldsymbol{X}\boldsymbol{C} - h\boldsymbol{X} \\ &= [\boldsymbol{C}, \boldsymbol{X}]. && \text{the scalar-shift terms cancel}\end{aligned}$$

Next, expanding both factors of the shifted mean gives

$$\begin{aligned}&\mathrm{tr}\big(\boldsymbol{X}(\boldsymbol{C}+h\boldsymbol{I})\boldsymbol{X}(\boldsymbol{C}+h\boldsymbol{I})\big) \\ &\quad= \mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) + h\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}) + h\,\mathrm{tr}(\boldsymbol{X}^2\boldsymbol{C}) + h^2\,\mathrm{tr}(\boldsymbol{X}^2) \\ &\quad= \mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) + 2h\,\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2) + h^2\,\mathrm{tr}(\boldsymbol{X}^2) \\ &\quad= \mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) + 2h\mu + h^2.\end{aligned}$$

The second equality uses cyclicity of the trace. The last substitutes $\mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2) = \mu$ and $\mathrm{tr}(\boldsymbol{X}^2) = 1$. The denominator changes according to

$$\begin{aligned}\mathrm{tr}\big((\boldsymbol{C}+h\boldsymbol{I})\boldsymbol{X}^2\big) &= \mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2) + h\,\mathrm{tr}(\boldsymbol{X}^2) \\ &= \mu + h.\end{aligned}$$

The original quotient satisfies

$$2\mu\, q_{\boldsymbol{C},b}(\boldsymbol{X}) = \mathbb{E}\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{\Delta}_t\boldsymbol{X}\boldsymbol{\Delta}_t) + c_b\,\mathrm{tr}(\boldsymbol{X}\boldsymbol{C}\boldsymbol{X}\boldsymbol{C}) - k_b\,\|[\boldsymbol{C}, \boldsymbol{X}]\|_F^2.$$

Thus the shifted numerator adds only $c_b(2h\mu + h^2)$ to this expression. Using the denominator computed above,

$$\begin{aligned}q_{\boldsymbol{C}+h\boldsymbol{I},b}(\boldsymbol{X}) &= \frac{2\mu\, q_{\boldsymbol{C},b}(\boldsymbol{X}) + c_b(2h\mu + h^2)}{2(\mu+h)} \\ &= \frac{\mu\, q_{\boldsymbol{C},b}(\boldsymbol{X}) + c_b h(\mu + h/2)}{\mu + h}.\end{aligned}$$

The last equality divides numerator and denominator by 2. This is (66) and completes the proof of Lemma G.3. □

Proof of Proposition G.2. Fix a PSD test matrix $\boldsymbol{X}$ with $\|\boldsymbol{X}\|_F = 1$, and recall $\mu = \mathrm{tr}(\boldsymbol{C}\boldsymbol{X}^2) > 0$. We divide the trajectories by a geometric sequence to apply the exact criterion, compare the resulting quotient with $q_{\boldsymbol{C},b}(\boldsymbol{X})$, and bound the difference by scalar algebra.

Step 1. Divide out enough growth to obtain stability.

By assumption, $\rho(\mathcal{T}) \le R$. Choose any real number $s > \sqrt{R} \ge 1$. Dividing the state at time t by s^t divides its one-step transition matrix by s and its second-moment map by s^2. Indeed, for the existing history state $\boldsymbol{y}_t = (\boldsymbol{x}_t, \boldsymbol{x}_{t-1})^\top$ and transition matrix $\boldsymbol{F}_t$,

$$s^{-(t+1)}\boldsymbol{y}_{t+1} = \frac{\boldsymbol{F}_t}{s}\,(s^{-t}\boldsymbol{y}_t),$$
$$\mathbb{E}\left[\frac{\boldsymbol{F}_t}{s}\boldsymbol{S}\frac{\boldsymbol{F}_t^\top}{s}\right] = \frac{1}{s^2}\mathbb{E}[\boldsymbol{F}_t\boldsymbol{S}\boldsymbol{F}_t^\top] = \frac{1}{s^2}\mathcal{T}(\boldsymbol{S}).$$

The last identity holds for every symmetric matrix $\boldsymbol{S}$. Hence the new moment map has spectral radius

$$\rho\left(\frac{1}{s^2}\mathcal{T}\right) = \frac{\rho(\mathcal{T})}{s^2} \le \frac{R}{s^2} < 1.$$

By the second-moment characterization (Lemma G.14), the divided state is strictly mean-square stable.

We now express this division in the position recurrence so that the exact criterion (Theorem G.1) applies. Set $\boldsymbol{x}'_t := s^{-t}\boldsymbol{x}_t$. Its history state satisfies

$$\begin{pmatrix}\boldsymbol{x}'_t\\ \boldsymbol{x}'_{t-1}\end{pmatrix} = \begin{pmatrix}\boldsymbol{I} & 0\\ 0 & s\boldsymbol{I}\end{pmatrix}(s^{-t}\boldsymbol{y}_t).$$

This fixed invertible change of coordinates preserves strict mean-square stability. Thus the recurrence for $\boldsymbol{x}'_t$ is strictly stable for every initial pair. To find that recurrence, substitute $\boldsymbol{x}_t = s^t\boldsymbol{x}'_t$ and $\boldsymbol{x}_{t-1} = s^{t-1}\boldsymbol{x}'_{t-1}$ into the original update:

$$\begin{aligned}\boldsymbol{x}'_{t+1} &= s^{-(t+1)}\left(\big[(1+b)\boldsymbol{I} - \alpha(\boldsymbol{C}+\boldsymbol{\Delta}_t)\big]\boldsymbol{x}_t - b\boldsymbol{x}_{t-1}\right)\\ &= s^{-(t+1)}\left(\big[(1+b)\boldsymbol{I} - \alpha(\boldsymbol{C}+\boldsymbol{\Delta}_t)\big]s^t\boldsymbol{x}'_t - bs^{t-1}\boldsymbol{x}'_{t-1}\right)\\ &= \left[\frac{1+b}{s}\boldsymbol{I} - \frac{\alpha}{s}(\boldsymbol{C}+\boldsymbol{\Delta}_t)\right]\boldsymbol{x}'_t - \frac{b}{s^2}\boldsymbol{x}'_{t-1}.\end{aligned}$$

The new momentum coefficient is b/s^2. The generic recurrence uses the identity coefficient $1 + b/s^2$, whereas the last display has $(1+b)/s$. We absorb their difference into a scalar shift of the mean. Define

$$\begin{aligned}u &:= \frac{s}{\alpha}\left(1 + \frac{b}{s^2} - \frac{1+b}{s}\right)\\ &= \frac{s^2 + b - s(1+b)}{s\alpha}\\ &= \frac{(s-1)(s-b)}{s\alpha} \ge 0.\end{aligned}$$

Here both factors in the numerator are positive because $s > 1 > b$. By construction,

$$\left(1 + \frac{b}{s^2}\right)\boldsymbol{I} - \frac{\alpha}{s}(\boldsymbol{C} + u\boldsymbol{I} + \boldsymbol{\Delta}_t) = \frac{1+b}{s}\boldsymbol{I} - \frac{\alpha}{s}(\boldsymbol{C}+\boldsymbol{\Delta}_t).$$

Consequently, the divided positions obey

$$\boldsymbol{x}'_{t+1} = \left[\left(1 + \frac{b}{s^2}\right)\boldsymbol{I} - \frac{\alpha}{s}(\boldsymbol{C} + u\boldsymbol{I} + \boldsymbol{\Delta}_t)\right]\boldsymbol{x}'_t - \frac{b}{s^2}\boldsymbol{x}'_{t-1}.$$

This is the generic recurrence with mean $\boldsymbol{C} + u\boldsymbol{I} \succ 0$, momentum $b/s^2 \in [0,1)$, coefficient $\alpha/s > 0$, and the same noise $\boldsymbol{\Delta}_t$. These parameters satisfy the assumptions of Theorem G.1.

Step 2. Compare the quotient at the same test matrix.

Keep the matrix $\boldsymbol{X}$ from the statement fixed. The divided recurrence is strictly stable, so the exact criterion (Theorem G.1) gives

$$\begin{aligned}\frac{s - b/s}{\alpha} &= \frac{1 - b/s^2}{\alpha/s} \\ &> r_{\boldsymbol{C}+u\boldsymbol{I}}(b/s^2) \\ &\geq q_{\boldsymbol{C}+u\boldsymbol{I},\, b/s^2}(\boldsymbol{X}).\end{aligned} \tag{110}$$

The last inequality holds because $\boldsymbol{X} \succeq 0$ and $\|\boldsymbol{X}\|_F = 1$ make it an admissible test matrix. We compare this last quotient with $q_{\boldsymbol{C},b}(\boldsymbol{X})$ by undoing the mean shift and comparing the momenta.

First, recall that $\mu = \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}^2)$ and $\operatorname{tr}(\boldsymbol{X}^2) = 1$, so

$$\operatorname{tr}\big((\boldsymbol{C} + u\boldsymbol{I})\boldsymbol{X}^2\big) = \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}^2) + u\operatorname{tr}(\boldsymbol{X}^2) = \mu + u.$$

The shift identity (Lemma G.3), with the same noise in both quotients, gives

$$q_{\boldsymbol{C}+u\boldsymbol{I},\, b/s^2}(\boldsymbol{X}) = \frac{\mu\, q_{\boldsymbol{C},b/s^2}(\boldsymbol{X}) + c_{b/s^2}u(\mu + u/2)}{\mu + u}.$$

Second, $b \geq 0$ and $s > 1$ imply $b/s^2 \leq b$. The pointwise momentum inequality (65) therefore gives

$$q_{\boldsymbol{C},b/s^2}(\boldsymbol{X}) \geq q_{\boldsymbol{C},b}(\boldsymbol{X}).$$

Substituting the shift identity into (110), multiplying by $\mu + u > 0$, and then applying this comparison,

$$\begin{aligned}\frac{(s - b/s)(\mu + u)}{\alpha} &\geq \mu\, q_{\boldsymbol{C},b/s^2}(\boldsymbol{X}) + c_{b/s^2}u(\mu + u/2) \\ &\geq \mu\, q_{\boldsymbol{C},b}(\boldsymbol{X}) + c_{b/s^2}u(\mu + u/2).\end{aligned} \tag{111}$$

We have weakened the strict first inequality to a non-strict one. The right side now contains the quotient of the original recurrence at the original momentum.

Step 3. Bound the quotient difference.

Moving all other terms in (111) to the right gives

$$\begin{aligned}\mu\left(q_{\boldsymbol{C},b}(\boldsymbol{X}) - \frac{1-b}{\alpha}\right) &\leq \frac{(s - b/s)(\mu + u)}{\alpha} - \frac{\mu(1-b)}{\alpha} - c_{b/s^2}u\left(\mu + \frac{u}{2}\right) \\ &= \mu\left[\frac{s - b/s - (1-b)}{\alpha} - c_{b/s^2}u\right] + u\left[\frac{s - b/s}{\alpha} - \frac{c_{b/s^2}u}{2}\right].\end{aligned}$$

The first numerator simplifies as

$$s - \frac{b}{s} - (1-b) = (s-1) + b\left(1 - \frac{1}{s}\right) = (s-1)\left(1 + \frac{b}{s}\right).$$

Dividing by $\mu > 0$, we obtain

$$q_{\boldsymbol{C},b}(\boldsymbol{X}) - \frac{1-b}{\alpha} \leq \left[\frac{(s-1)(1+b/s)}{\alpha} - c_{b/s^2}u\right] + \frac{u}{\mu}\left[\frac{s - b/s}{\alpha} - \frac{c_{b/s^2}u}{2}\right].$$

We simplify these two terms separately. Recall the scalar shift from Step 1 and the definition of c_b:

$$u = \frac{(s-1)(s-b)}{s\alpha}, \qquad c_{b/s^2} = \frac{1 - b/s^2}{1 + b/s^2} = \frac{s^2 - b}{s^2 + b}.$$

For the first term, substituting these expressions and combining denominators gives

$$\begin{aligned}\frac{(s-1)(1+b/s)}{\alpha} - c_{b/s^2}u &= \frac{s-1}{s\alpha}\left[(s+b) - \frac{(s^2-b)(s-b)}{s^2+b}\right] \\ &= \frac{(s-1)\big[(s+b)(s^2+b) - (s^2-b)(s-b)\big]}{s\alpha(s^2+b)} \\ &= \frac{(s-1)(2bs^2 + 2bs)}{s\alpha(s^2+b)} \\ &= \frac{2b(s-1)(s+1)}{\alpha(s^2+b)} = \frac{2b(s^2-1)}{(s^2+b)\alpha}.\end{aligned}$$

For the second term, first substitute u inside the bracket, then factor its numerator:

$$\begin{aligned}
&\frac{u}{\mu}\left[\frac{s-b/s}{\alpha}-\frac{c_{b/s^2}u}{2}\right] \\
&\quad=\frac{u(s^2-b)}{\mu s\alpha}\left[1-\frac{(s-1)(s-b)}{2(s^2+b)}\right] \\
&\quad=\frac{u(s^2-b)\big[2(s^2+b)-(s-1)(s-b)\big]}{2\mu s\alpha(s^2+b)} \\
&\quad=\frac{u(s^2-b)\big[s^2+(b+1)s+b\big]}{2\mu s\alpha(s^2+b)} \\
&\quad=\frac{u(s^2-b)(s+1)(s+b)}{2\mu s\alpha(s^2+b)} \\
&\quad=\frac{(s-1)(s-b)(s^2-b)(s+1)(s+b)}{2\mu s^2\alpha^2(s^2+b)} \\
&\quad=\frac{(s^2-1)(s^2-b^2)(s^2-b)}{2s^2(s^2+b)\mu\alpha^2}.
\end{aligned}$$

The penultimate line substitutes the remaining factor u. The last line groups $(s-1)(s+1)$ and $(s-b)(s+b)$. Adding the two terms yields, for every $s>\sqrt{R}$,

$$q_{\boldsymbol{C},b}(\boldsymbol{X})-\frac{1-b}{\alpha}\le(s^2-1)\left[\frac{2b}{(s^2+b)\alpha}+\frac{(s^2-b^2)(s^2-b)}{2s^2(s^2+b)\mu\alpha^2}\right].$$

The left side does not depend on s, and the right side is continuous at $s=\sqrt{R}$, with positive denominators. Letting $s\downarrow\sqrt{R}$ therefore gives

$$q_{\boldsymbol{C},b}(\boldsymbol{X})-\frac{1-b}{\alpha}\le(R-1)\left[\frac{2b}{(R+b)\alpha}+\frac{(R-b^2)(R-b)}{2R(R+b)\mu\alpha^2}\right]. \tag{112}$$

We retain this sharper form and now simplify its second coefficient.

Step 4. Simplify the second coefficient.

The second coefficient in (112) is easier to compare after bounding it by a square. To justify this bound, subtract its two relevant factors:

$$\begin{aligned}
\frac{R-b}{R}-\frac{R-b^2}{R+b}&=\frac{(R-b)(R+b)-R(R-b^2)}{R(R+b)} \\
&=\frac{R^2-b^2-R^2+Rb^2}{R(R+b)} \\
&=\frac{b^2(R-1)}{R(R+b)}\ge 0.
\end{aligned}$$

Multiplying by $(R-b)/R>0$ gives

$$\frac{(R-b^2)(R-b)}{R(R+b)}\le\left(\frac{R-b}{R}\right)^2=\left(1-\frac{b}{R}\right)^2.$$

Thus (112) implies

$$q_{\boldsymbol{C},b}(\boldsymbol{X})-\frac{1-b}{\alpha}\le(R-1)\left[\frac{2b}{(R+b)\alpha}+\frac{1}{2\mu}\left(\frac{1-b/R}{\alpha}\right)^2\right].$$

This is (64) and completes the proof of Proposition G.2. □

G.5.3 Proof of the AGM Critical LR

Recall from Appendix G.1 that $b = \beta_1(1 - \eta\lambda_{\mathrm{WD}})$, $\alpha = \eta(1 - \beta_1)$, and

$$G_{\boldsymbol{C}}(\eta) = \frac{1}{\eta} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1 - \beta_1} - r_{\boldsymbol{C}}\big(\beta_1(1 - \eta\lambda_{\mathrm{WD}})\big).$$

We prove that this function has one zero, which is the endpoint of the stable LRs, and that the moment spectral radius is strictly greater than one above it.

Proof of Corollary G.4. To rule out $\rho(\mathcal{T}) \le 1$ above the root, we rescale a trajectory by s^{-t} and let $s \downarrow 1$.

Step 1. Apply the fixed-parameter criterion.

For $\eta < \eta_{\max}$, Theorem G.1 gives strict stability if and only if $G_{\boldsymbol{C}}(\eta) > 0$. If $\eta_{\max} < \infty$ and $\eta \ge \eta_{\max}$, then $b \le -1$. The roots of the mean recurrence along each eigendirection have product b, so one root has modulus at least one. Strict mean square stability then fails by the mean-radius inequality (113). At $\eta\lambda_{\mathrm{WD}} = 1$, the effective momentum is simply $b = 0$, already covered by the criterion.

Step 2. Prove monotonicity and continuity.

Recall the pointwise momentum comparison (65): for $-1 < b' \le b < 1$, $q_{\boldsymbol{C},b'}(\boldsymbol{X}) \ge q_{\boldsymbol{C},b}(\boldsymbol{X})$ for every nonzero PSD $\boldsymbol{X}$. We apply this comparison before taking maxima.

The maximization defining $r_{\boldsymbol{C}}$ uses the same set $\{\boldsymbol{X} \succeq 0 : \|\boldsymbol{X}\|_F = 1\}$ for every b. Choose a maximizer $\boldsymbol{X}_b$ at b, which exists by (50). It satisfies $\boldsymbol{X}_b \succeq 0$, $\|\boldsymbol{X}_b\|_F = 1$, and $q_{\boldsymbol{C},b}(\boldsymbol{X}_b) = r_{\boldsymbol{C}}(b)$. Therefore,

$$\begin{aligned}
r_{\boldsymbol{C}}(b') &= \max_{\boldsymbol{X}\succeq 0,\ \|\boldsymbol{X}\|_F = 1} q_{\boldsymbol{C},b'}(\boldsymbol{X}) && \text{definition of } r_{\boldsymbol{C}}(b') \\
&\ge q_{\boldsymbol{C},b'}(\boldsymbol{X}_b) && \boldsymbol{X}_b \text{ is also admissible at } b' \\
&\ge q_{\boldsymbol{C},b}(\boldsymbol{X}_b) && \text{the pointwise inequality above} \\
&= r_{\boldsymbol{C}}(b). && \boldsymbol{X}_b \text{ maximizes the quotient at } b
\end{aligned}$$

Hence $b' \le b$ implies $r_{\boldsymbol{C}}(b') \ge r_{\boldsymbol{C}}(b)$, which is the claimed nonincreasing dependence on b.

Return now to the AGM parameter curve and set $b_i := \beta_1(1 - \eta_i\lambda_{\mathrm{WD}})$. Since $\eta_2 < \eta_1$, we have $b_2 \ge b_1$ and therefore $r_{\boldsymbol{C}}(b_1) \ge r_{\boldsymbol{C}}(b_2)$. Recall that $G_{\boldsymbol{C}}(\eta)$ subtracts $r_{\boldsymbol{C}}(b)$ from $1/\eta + \beta_1\lambda_{\mathrm{WD}}/(1 - \beta_1)$. The offset $\beta_1\lambda_{\mathrm{WD}}/(1 - \beta_1)$ is constant along this curve, giving

$$\begin{aligned}
G_{\boldsymbol{C}}(\eta_2) - G_{\boldsymbol{C}}(\eta_1) &= \frac{1}{\eta_2} - \frac{1}{\eta_1} + r_{\boldsymbol{C}}(b_1) - r_{\boldsymbol{C}}(b_2) \\
&\ge \frac{1}{\eta_2} - \frac{1}{\eta_1} > 0. && \text{monotonicity of } r_C
\end{aligned}$$

This is (71). The quotient is jointly continuous in $b \in (-1, 1)$ and $\boldsymbol{X}$ on the compact PSD unit sphere, with denominator at least $2\lambda_{\min}(\boldsymbol{C})$. Its maximum $r_{\boldsymbol{C}}(b)$, and hence $G_{\boldsymbol{C}}$, is continuous.

Step 3. Check the endpoints.

As $\eta \downarrow 0$, $b \to \beta_1 < 1$ and

$$G_{\boldsymbol{C}}(\eta) = \frac{1}{\eta} + \frac{\beta_1\lambda_{\mathrm{WD}}}{1 - \beta_1} - r_{\boldsymbol{C}}(b) \longrightarrow +\infty.$$

If $\beta_1\lambda_{\mathrm{WD}} > 0$, then $b \downarrow -1$ as $\eta \uparrow \eta_{\max}$. The rank-one lower bound gives

$$r_{\boldsymbol{C}}(b) \ge \frac{1 - b}{1 + b}\frac{L_{\boldsymbol{C}}}{2} \longrightarrow \infty, \qquad G_{\boldsymbol{C}}(\eta) \longrightarrow -\infty.$$

If $\beta_1\lambda_{\mathrm{WD}} = 0$, then $b = \beta_1$ is fixed and

$$G_{\boldsymbol{C}}(\eta) = \frac{1}{\eta} - r_{\boldsymbol{C}}(\beta_1) \longrightarrow -r_{\boldsymbol{C}}(\beta_1) < 0 \qquad (\eta \to \infty).$$

Continuity and strict monotonicity now give one root, with $G_{\boldsymbol{C}} > 0$ precisely below it.

Step 4. Locate the second-moment spectral radius.

First fix $\eta \in (\eta_{\text{crit}}, \eta_{\max})$, so that $-1 < b < 1$ and $G_{\boldsymbol{C}}(\eta) < 0$. Suppose for a contradiction that its second-moment map satisfies $\rho(\mathcal{T}) \le 1$. For every $s > 1$, dividing the state at time t by s^t gives

$$\rho(s^{-2}\mathcal{T}) = s^{-2}\rho(\mathcal{T}) \le s^{-2} < 1.$$

Recall the position calculation in the proof of Proposition G.2. The divided recurrence has coefficient α/s, momentum b/s^2, unchanged noise, and mean $\boldsymbol{C} + u\boldsymbol{I}$, where

$$u := \frac{(s-1)(s-b)}{s\alpha} > 0, \qquad 1 + \frac{b}{s^2} - \frac{\alpha}{s}u = \frac{1+b}{s}.$$

That calculation needs only $-1 < b < 1$: for every $s > 1$ we have $-1 < b/s^2 < 1$, $\alpha/s > 0$, and $u > 0$ because $s > b$. Thus the shifted mean remains positive definite even when $b < 0$. The divided recurrence is strictly stable, so the exact criterion gives, for every $\boldsymbol{X} \succeq 0$ with $\|\boldsymbol{X}\|_F = 1$,

$$\begin{aligned} q_{\boldsymbol{C}+u\boldsymbol{I},\, b/s^2}(\boldsymbol{X}) &\le r_{\boldsymbol{C}+u\boldsymbol{I}}(b/s^2) \\ &< \frac{1 - b/s^2}{\alpha/s} = \frac{s - b/s}{\alpha}. \end{aligned}$$

As $s \downarrow 1$, we have $u \to 0$ and $b/s^2 \to b$. The quotient is continuous in these parameters, with a positive denominator. Hence

$$\begin{aligned} q_{\boldsymbol{C},b}(\boldsymbol{X}) &\le \frac{1-b}{\alpha} && \text{take the limit for each fixed } \boldsymbol{X} \\ r_{\boldsymbol{C}}(b) &= \max_{\boldsymbol{X}\succeq 0,\ \|\boldsymbol{X}\|_F = 1} q_{\boldsymbol{C},b}(\boldsymbol{X}) \le \frac{1-b}{\alpha}, \\ G_{\boldsymbol{C}}(\eta) &= \frac{1-b}{\alpha} - r_{\boldsymbol{C}}(b) \ge 0. \end{aligned}$$

This contradicts $G_{\boldsymbol{C}}(\eta) < 0$. Thus $\rho(\mathcal{T}) > 1$ throughout this interval.

If $\eta_{\max} < \infty$ and $\eta > \eta_{\max}$, then $b < -1$. The two roots along any eigendirection of the mean have product b, so at least one has modulus greater than one. At $\eta = \eta_{\max}$, we have $b = -1$. For an eigenvalue $a_i > 0$ of $\boldsymbol{C}$, the characteristic polynomial satisfies

$$p_i(x) = x^2 + \alpha a_i x - 1, \qquad p_i(-1) = -\alpha a_i < 0, \qquad \lim_{x\to-\infty} p_i(x) = +\infty.$$

It therefore has a root below -1. In both cases, the mean-radius inequality (113) implies $\rho(\mathcal{T}) > 1$.

Finally, the entries of the second-moment map are continuous in η, so its spectral radius is continuous. The stable rates approaching η_{crit} from below give $\rho(\mathcal{T}) \le 1$ at the critical rate. Since $G_{\boldsymbol{C}}(\eta_{\text{crit}}) = 0$, the exact criterion excludes strict stability there and gives $\rho(\mathcal{T}) \ge 1$. Thus the radius equals one at the critical rate. This completes the proof of Corollary G.4. □

G.6 Auxiliary Results

The proofs use four finite-dimensional lemmas on second-moment maps and on maps that preserve positive semidefinite matrices. We also use the Cauchy–Schwarz and Minkowski inequalities, Jensen's inequality, the Neumann series for $(s\mathcal{I} - \mathcal{U})^{-1}$ when $|s| > \rho(\mathcal{U})$, the Rayleigh quotient characterization of the largest eigenvalue of a self-adjoint operator, and continuity of polynomial roots in their coefficients. The cutoff argument uses the Jordan decomposition: the powers of a finite-dimensional operator with spectral radius one grow at most polynomially.

We distinguish two notions of positivity. A map $\mathcal{U} : \mathbb{S}_n \to \mathbb{S}_n$ *preserves PSD matrices* if $\boldsymbol{X} \succeq 0$ implies $\mathcal{U}(\boldsymbol{X}) \succeq 0$. A self-adjoint operator $\mathcal{D}$ on $\mathbb{S}_n$ is *positive definite*, written $\mathcal{D} \succ 0$, if $\langle \boldsymbol{X}, \mathcal{D}\boldsymbol{X}\rangle > 0$ for every nonzero symmetric $\boldsymbol{X}$. Preservation of PSD matrices does not imply positive definiteness as an operator, and the converse fails as well. Second moments are uncentered throughout: $\boldsymbol{\Sigma}_t = \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}_t^\top]$, so that $\operatorname{tr} \boldsymbol{\Sigma}_t = \mathbb{E}\,\|\boldsymbol{y}_t\|^2$, and no zero-mean assumption is imposed on the state.

Lemma G.14 (Second-moment characterization). *Let $\boldsymbol{y}_{t+1} = \boldsymbol{M}_t \boldsymbol{y}_t$, where the $\boldsymbol{M}_t$ are i.i.d. with finite second moments and $\boldsymbol{y}_0$ is deterministic. Define $\mathcal{T}(\boldsymbol{S}) := \mathbb{E}[\boldsymbol{M}_t \boldsymbol{S} \boldsymbol{M}_t^\top]$ on $\mathbb{S}_n$. Then $\mathbb{E}\,\|\boldsymbol{y}_t\|^2 \to 0$ for every initial state if and only if $\rho(\mathcal{T}) < 1$. Strict mean-square stability also implies $\rho(\mathbb{E}\boldsymbol{M}_t) < 1$, and in general*

$$\rho(\mathbb{E}\boldsymbol{M}_t)^2 \le \rho(\mathcal{T}). \tag{113}$$

Proof of Lemma G.14. Write $\boldsymbol{\Sigma}_t := \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}_t^\top]$. Since $\boldsymbol{y}_t$ depends only on the initial state and on $\boldsymbol{M}_0, \dots, \boldsymbol{M}_{t-1}$, it is independent of $\boldsymbol{M}_t$, so

$$\begin{aligned}
\boldsymbol{\Sigma}_{t+1} &= \mathbb{E}[\boldsymbol{M}_t \boldsymbol{y}_t \boldsymbol{y}_t^\top \boldsymbol{M}_t^\top] \\
&= \mathbb{E}\big[\boldsymbol{M}_t\, \mathbb{E}[\boldsymbol{y}_t \boldsymbol{y}_t^\top \mid \boldsymbol{M}_t]\, \boldsymbol{M}_t^\top\big] \\
&= \mathbb{E}[\boldsymbol{M}_t \boldsymbol{\Sigma}_t \boldsymbol{M}_t^\top] = \mathcal{T}(\boldsymbol{\Sigma}_t),
\end{aligned}$$

with $\boldsymbol{\Sigma}_0 = \boldsymbol{y}_0 \boldsymbol{y}_0^\top$. The second equality uses the tower property, conditioning on $\boldsymbol{M}_t$, and the third uses independence. If $\rho(\mathcal{T}) < 1$, the Jordan form of $\mathcal{T}$ shows that $\|\mathcal{T}^t\|$ is bounded by a polynomial in t times a decaying exponential, so $\mathcal{T}^t \to 0$ and $\operatorname{tr} \boldsymbol{\Sigma}_t \to 0$.

Conversely, suppose $\mathbb{E}\,\|\boldsymbol{y}_t\|^2 \to 0$ for every $\boldsymbol{y}_0$. Since $\boldsymbol{\Sigma}_t \succeq 0$ we have $\|\boldsymbol{\Sigma}_t\|_F \le \operatorname{tr} \boldsymbol{\Sigma}_t$, so $\mathcal{T}^t(\boldsymbol{y}_0 \boldsymbol{y}_0^\top) = \boldsymbol{\Sigma}_t \to 0$ for every $\boldsymbol{y}_0$. Rank-one symmetric matrices span $\mathbb{S}_n$: the diagonal basis elements are $\boldsymbol{e}_i \boldsymbol{e}_i^\top$, and each off-diagonal symmetric basis element is $(\boldsymbol{e}_i + \boldsymbol{e}_j)(\boldsymbol{e}_i + \boldsymbol{e}_j)^\top - \boldsymbol{e}_i \boldsymbol{e}_i^\top - \boldsymbol{e}_j \boldsymbol{e}_j^\top$. Hence $\mathcal{T}^t \to 0$ as an operator on the finite-dimensional space $\mathbb{S}_n$. An eigenvalue of modulus at least one would contradict this convergence, so $\rho(\mathcal{T}) < 1$.

For the statements about the mean, independence gives $\mathbb{E}\boldsymbol{y}_{t+1} = (\mathbb{E}\boldsymbol{M}_t)\mathbb{E}\boldsymbol{y}_t$, and the matrices have the same law at each time. Iterating and applying Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\boldsymbol{y}_t &= (\mathbb{E}\boldsymbol{M}_t)^t \boldsymbol{y}_0, \\
\big\|(\mathbb{E}\boldsymbol{M}_t)^t \boldsymbol{y}_0\big\|^2 &= \|\mathbb{E}\boldsymbol{y}_t\|^2 \le \mathbb{E}\,\|\boldsymbol{y}_t\|^2 = \operatorname{tr} \mathcal{T}^t(\boldsymbol{y}_0 \boldsymbol{y}_0^\top).
\end{aligned}$$

Under strict mean-square stability the right side tends to zero for every $\boldsymbol{y}_0$, so $(\mathbb{E}\boldsymbol{M}_t)^t \to 0$ and $\rho(\mathbb{E}\boldsymbol{M}_t) < 1$. To obtain (113), apply this implication to the matrices $\boldsymbol{M}_t/\sqrt{s}$ for any real $s > \rho(\mathcal{T})$. Their second-moment map is $\mathcal{T}/s$, so

$$\begin{aligned}
\rho(\mathcal{T}/s) = \rho(\mathcal{T})/s < 1, &\qquad \text{choice of } s \\
\rho\big(\mathbb{E}\boldsymbol{M}_t/\sqrt{s}\big) < 1, &\qquad \text{the mean-stability implication above} \\
\rho(\mathbb{E}\boldsymbol{M}_t)^2 < s. &\qquad \text{rescale the mean matrix}
\end{aligned}$$

Letting $s \downarrow \rho(\mathcal{T})$ proves (113), including when $\rho(\mathcal{T}) = 0$. This completes the proof of Lemma G.14. □

Connection to the Kronecker second-moment matrix. The matrix $\mathbb{E}[\boldsymbol{M}_t \otimes \boldsymbol{M}_t]$ represents the map $\boldsymbol{S} \mapsto \mathbb{E}[\boldsymbol{M}_t \boldsymbol{S} \boldsymbol{M}_t^\top]$ on all $n \times n$ matrices in the column-stacked basis. For the projected AGM dynamics, this is $\boldsymbol{K}^+(\eta)$ in (18). The map $\mathcal{T}$ in the preceding lemma is the restriction to symmetric matrices.

Write $\boldsymbol{\Phi}_t := \boldsymbol{M}_{t-1} \cdots \boldsymbol{M}_0$. For fixed vectors $\boldsymbol{u}, \boldsymbol{v}$, the Cauchy–Schwarz inequality gives

$$\begin{aligned}
\big\|\mathbb{E}[\boldsymbol{\Phi}_t \boldsymbol{u} \boldsymbol{v}^\top \boldsymbol{\Phi}_t^\top]\big\|_F &\le \mathbb{E}\,\big\|\boldsymbol{\Phi}_t \boldsymbol{u} \boldsymbol{v}^\top \boldsymbol{\Phi}_t^\top\big\|_F \\
&= \mathbb{E}\big[\|\boldsymbol{\Phi}_t \boldsymbol{u}\|\,\|\boldsymbol{\Phi}_t \boldsymbol{v}\|\big] \\
&\le \big(\mathbb{E}\,\|\boldsymbol{\Phi}_t \boldsymbol{u}\|^2\ \mathbb{E}\,\|\boldsymbol{\Phi}_t \boldsymbol{v}\|^2\big)^{1/2}.
\end{aligned}$$

If the symmetric second moments decay, the right side tends to zero. Since the matrices $\boldsymbol{u}\boldsymbol{v}^\top$ span all matrices, the full moment map also has powers tending to zero. The converse follows by restriction. Replacing $\boldsymbol{M}_t$ by $\boldsymbol{M}_t/\sqrt{s}$ divides both moment maps by s. Applying the same argument for every $s > 0$ shows that the two spectral radii are below s for exactly the same s, hence are equal.

Recall that the history state uses $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$, whereas (67) uses position and momentum $(\boldsymbol{x}_t, \boldsymbol{n}_t)$. The position update gives their linear relation

$$\begin{pmatrix} \boldsymbol{x}_t \\ \boldsymbol{n}_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & 0 \\ -\eta^{-1}\boldsymbol{I} & \eta^{-1}(1 - \eta\lambda_{\mathrm{WD}})\boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{pmatrix}.$$

This change is invertible when $\eta\lambda_{\mathrm{WD}} \neq 1$, and an invertible change of state induces a similarity between the moment maps. Their spectral radii are therefore equal there. The entries of both moment maps are polynomials in η, and the spectral radius is continuous in the entries, so the equality extends to $\eta\lambda_{\mathrm{WD}} = 1$ as well.

Lemma G.15 (A Lyapunov criterion for PSD-preserving maps)**.** *If $\mathcal{U} : \mathbb{S}_n \to \mathbb{S}_n$ preserves PSD matrices, then*

$$\rho(\mathcal{U}) < 1 \quad \Longleftrightarrow \quad \text{there exists } \boldsymbol{W} \succ 0 \text{ with } \boldsymbol{W} - \mathcal{U}(\boldsymbol{W}) \succ 0.$$

Proof of Lemma G.15. If $\rho(\mathcal{U}) < 1$, the Neumann series $\boldsymbol{W} := \sum_{j\geq 0} \mathcal{U}^j(\boldsymbol{I})$ converges, satisfies $\boldsymbol{W} \succeq \boldsymbol{I}$ because every summand is PSD, and $\boldsymbol{W} - \mathcal{U}(\boldsymbol{W}) = \boldsymbol{I}$. Conversely, given such a $\boldsymbol{W}$, choose $\varepsilon \in (0, 1)$ with $0 \preceq \mathcal{U}(\boldsymbol{W}) \preceq (1-\varepsilon)\boldsymbol{W}$. For every symmetric $\boldsymbol{X}$ there is a finite a with $-a\boldsymbol{W} \preceq \boldsymbol{X} \preceq a\boldsymbol{W}$. Since $\mathcal{U}$ preserves the order $\preceq$,

$$-a(1-\varepsilon)^j \boldsymbol{W} \preceq \mathcal{U}^j(\boldsymbol{X}) \preceq a(1-\varepsilon)^j \boldsymbol{W} \qquad (j \geq 0).$$

Applying this to a finite basis of $\mathbb{S}_n$ gives $\mathcal{U}^j \to 0$, hence $\rho(\mathcal{U}) < 1$. This completes the proof of Lemma G.15. □

Lemma G.16 (Perron–Frobenius theorem for PSD-preserving maps)**.** *Let $\mathcal{U} : \mathbb{S}_n \to \mathbb{S}_n$ preserve PSD matrices, and suppose $R = \rho(\mathcal{U}) > 0$. There is a nonzero $\boldsymbol{W} \in \mathbb{S}_n^+$ with $\mathcal{U}(\boldsymbol{W}) = R\boldsymbol{W}$.*

This is the Perron–Frobenius theorem for a linear map that leaves a proper cone invariant, here the cone $\mathbb{S}_n^+$ of positive semidefinite matrices (Vandergraft, 1968; Berman and Plemmons, 1994, Chapter 1).

Lemma G.17 (Jury stability test for a real quadratic)**.** *Both roots of $p(x) = x^2 - \sigma x + \pi$, with real σ, π, lie strictly inside the unit disk if and only if $|\pi| < 1$, $1 - \sigma + \pi > 0$, and $1 + \sigma + \pi > 0$.*

This is the Schur–Cohn, or Jury, stability test for a real quadratic (Jury, 1964): the conditions are $|\pi| < 1$, $p(1) > 0$ and $p(-1) > 0$, and they are equivalent to both roots lying in the open unit disk.